Improving Caching Efficiency in Content-Aware C-RAN-Based Cooperative Beamforming: A Joint Design Approach

Ming-Min Zhao, Yunlong Cai, *Senior Member, IEEE*, Min-Jian Zhao, *Member, IEEE*, Benoit Champagne, *Senior Member, IEEE*, and Theodoros A. Tsiftsis, *Senior Member, IEEE*

***Abstract*— This work studies the joint problem of content placement, remote radio head (RRH) clustering and beamformer design, in a cache-enabled cloud-radio access network (C-RAN). In the considered system, downlink users are cooperatively served by multiple RRHs, in turn connected to a centralized baseband unit (BBU) pool via fronthaul links. Each RRH is equipped with a local cache from which it can directly acquire the requested user contents, without utilizing the fronthaul links. We aim to jointly optimize the aforementioned three aspects, in order to strike a balance between fronthaul traffic reduction and transmission power minimization. To this end, we propose to employ the ratio between these two important system utilities as the objective function, referred to as *caching efficiency*. Two joint design algorithms are presented to address the resulting nonconvex optimization problem, which features coupling constraints and mixed-integer variables, namely: the penalty concave-convex procedure (P-CCCP) and penalty dual decomposition (PDD) based algorithms. Furthermore, since content placement is usually updated over a larger timescale, we propose a two-timescale joint design algorithm, where the P-CCCP and PDD-based algorithms can be employed for efficient initialization as well as for establishing performance limits. Simulation results validate the efficiency of the proposed algorithms.**

***Index Terms*— Beamforming, caching, cloud-RAN, content placement, transceiver design, user-centric clustering.**

Manuscript received May 8, 2019; revised October 12, 2019 and December 23, 2019; accepted March 5, 2020. Date of publication March 18, 2020; date of current version June 10, 2020. The work of Ming-Min Zhao was supported in part by the National Natural Science Foundation of China under Grant 91938202, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LQ20F010010, and in part by the Fundamental Research Funds for the Central Universities under Grant 2019QNA5011. The work of Yunlong Cai was supported in part by the National Natural Science Foundation of China under Grant 61831004 and Grant 61971376 and in part by the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under Grant LR19F010002. The work of Min-Jian Zhao was supported by the National Key Research and Development Project under Grant 2018YFB1802303. This article was presented in part at the IEEE ICC 2019. The associate editor coordinating the review of this article and approving it for publication was A. Maaref. *(Corresponding author: Min-Jian Zhao.)*

Ming-Min Zhao, Yunlong Cai, and Min-Jian Zhao are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: zmmblack@zju.edu.cn; ylcai@zju.edu.cn; mjzho@zju.edu.cn).

Benoit Champagne is with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0E9, Canada (e-mail: benoit.champagne@mcgill.ca).

Theodoros A. Tsiftsis is with the School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai 519070, China (e-mail: theo_tsiftsis@jnu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2020.2979958

I. Introduction

With the increasing demand for high-speed data traffic, especially content sharing and video streaming, wireless network operators are faced with formidable challenges in attempting to provide high throughput and low latency services to large populations of mobile users. To meet these new service requirements, local caching of popular data at base stations (BSs) has been proposed recently as a promising solution for massive content delivery [2]–[8]. This approach essentially brings key information contents closer to the users, which in turn reduces fronthaul utilization costs and also eliminates a significant amount of backhaul traffic. Furthermore, as service providers move favored contents to intermediate nodes in the network, the access delay is reduced which improves the quality of experience for users.

A. Technical Literature Review

To support the ever increasing data traffic and computational demands of users, another key technology is that of cloud radio access network (C-RAN), which refers to an emerging network architecture that can improve the spectrum and energy efficiency of current wireless networks [9]–[12]. In C-RAN, several low-cost low-power remote radio heads (RRHs) are deployed to replace the traditional high-cost BSs. Since most of the signal processing tasks are handled by a centralized baseband unit (BBU) pool that connects to the RRHs via digital fronthaul links,[1] joint data processing and precoding are possible to improve system performance.

With regard to fronthaul/backhaul traffic reduction, RRH clustering and cooperative beamforming are attractive approaches since the popular data of each user only need to be assigned to a small cluster of serving RRHs, instead of all RRHs. In the literature, several recent works have investigated

[1]Note that in the C-RAN context, the backhaul portion of the network comprises the intermediate links between the core network and the BBU pool, while the links between the BBUs and the RRHs at the edge of the network are usually referred to as fronthaul links. In general, content caching at the RRHs would save both fronthaul and backhaul costs. However, for conciseness, we will only mention fronthaul in the following.

the dynamic BS clustering and beamforming problem [10], [11], [13], [14]. Specifically, in [10], the authors formulated the problem from a sparse optimization perspective, and proposed an efficient algorithm based on iteratively solving a sequence of group least absolute shrinkage and selection operator (LASSO) problems. In [11], a group sparse beamformer was proposed by adopting the weighted l_1/l_2-norm minimization to induce sparsity. A network utility maximization problem with per-BS backhaul capacity constraints was considered in [13]. In [14], the joint clustering and beamforming problem was formulated as a mixed integer nonlinear program, which was addressed by relaxing the discrete clustering function with a continuous exponential. Note that these works took a user-centric or connection-centric point of view for the joint design, and as such the impact of content placement was not considered.

As previously indicated, to further improve the delivery rate and decrease fronthaul costs and latency for mobile users, a promising solution is to cache popular contents at the RRHs. This content delivery service can be conducted by carefully designing content placement such that the users can seize various transmission opportunities and fully exploit the caching gain [15]. The potential benefits of distributing and storing popular contents across the whole network have been investigated by many researchers, as discussed below.

In [3], the authors introduced the concept of *FemtoCaching* in order to increase the throughput of wireless video delivery networks, and further investigated the wireless distributed caching problem with the aim to minimize the total average delay of all users. The NP-hardness of the caching problem was proved and numerical algorithms for its approximate solution were presented. The problem of content placement in Fog-RANs was considered in [16], by taking into account different file preferences and diverse transmission opportunities for each user. The joint optimization of data placement and beamforming vectors in backhaul limited networks was investigated in [17] by establishing the connection between data assignment and sparsity-inducing norm. In [18] and [19], it was reported that three distinct benefits can be achieved from caching, that is: load balancing gain, interference cancellation gain, and interference alignment gain. Distributed caching algorithms based on belief propagation were developed in [15], [20], [21] in order to minimize the downloading latency. The tradeoffs between small BS density and total cache size were investigated in [7], where it was shown that significant gains in terms of outage probability and average delivery rate are possible by employing cache-enabled small BSs. The joint design of content-centric BS clustering and multicast beamforming in the cache-enabled cloud RAN was investigated in [22] by assuming that the content placement is fixed according to a given caching strategy. Joint optimization of user association and content placement was investigated in [23], where the aim was to reduce the backhaul traffic in a densely deployed wireless access network. In [24], the interplay between cloud processing and edge caching was addressed from an information-theoretic viewpoint.

B. Motivation

While making significant advances, the aforementioned studies do not approach the problem of content placement, RRH clustering and downlink beamformer design by considering all three aspects jointly. In this work, we study the joint optimization of a generic content-aware C-RAN along these three critical design dimensions, aiming to strike a more favorable balance between the reduction of fronthaul traffic and transmission power. To this end, we shall seek to maximize the ratio between these two important system utilities, termed *caching efficiency*, subject to quality of service (QoS), clustering and caching constraints.[2] The joint design problem is quite challenging due to the facts that the fractional objective function and the constraints are both nonconvex, the optimization variables are tightly coupled and the latter contain nontrivial discrete variables. By exploiting the problem structure, we propose two algorithms which both take advantage of the Dinkelbach method [27]. To derive the first algorithm, we first transform the discrete constraints into alternative inequality constraints; then, by combining the benefits of the penalty method [28] and the concave-convex procedure (CCCP) [29], the transformed problem can be efficiently solved. This approach leads to a so-called penalty CCCP (P-CCCP) algorithm[3] which iterates over two steps, i.e.: approximately solving the penalized subproblem and updating the penalty parameter and Dinkelbach variable. To derive the second algorithm, we utilize the penalty dual decomposition (PDD) framework [31] and show that by carefully introducing auxiliary variables, the joint design problem can be tackled by iterating over a sequence of simple and efficient updates in the individual design variables. While these two algorithms exhibit similar performance in simulations, each one offers different advantages. In particular, the P-CCCP algorithm can converge in fewer iterations, while the PDD-based algorithm admits a simpler implementation.

Furthermore, since content placement is typically updated over a larger timescale than RRH clustering and beamforming, we propose a two-timescale joint design algorithm, which is based on the two-stage online successive convex approximation (TOSCA) framework in [32]. The aforementioned P-CCCP and PDD-based algorithms can be employed within this context as powerful initialization methods, while they can also be modified without any difficulty to solve the underlying short-term subproblems.

[2]Note that the definition of caching efficiency here is different from [25] and [26], where the cache hit ratio, the offloading gain and degrees of freedom gain per unit cache size are interpreted as measures of caching efficiency. The main motivation to take caching efficiency as the objective function is based on the observation that with increased transmission power, larger serving clusters can be formed, which further reduces the fronthaul utilization.

[3]While a general framework for the P-CCCP algorithm was presented in [30], our work is very different from the latter due to the following facts: 1) the system model is very different, which leads to distinct optimization problems; 2) in the proposed algorithm, we handle discrete constraints by transforming them into inequality constraints, while [30] does not involve discrete variables; 3) we integrate the Dinkelbach method into the P-CCCP algorithm under the proposed framework.

C. Our Contributions

The main contributions of this work can be summarized as follows:

1) A general optimization framework for content-aware transceiver design in cache-enabled C-RAN systems is proposed. The ratio between the fronthaul traffic reduction and transmission power, referred to as caching efficiency, is proposed as a new objective function, which is jointly optimized along the three critical design dimensions of content placement, RRH association and downlink beamforming.

2) Two joint design algorithms, i.e., the P-CCCP and the PDD-based algorithms, are proposed to address the resulting highly nonconvex problem with mixed-integer variables. A detailed complexity analysis of these two algorithms and their respective advantages and limitations are exposed. Furthermore, a two-timescale joint design algorithm is proposed, where the content placement is updated over a larger time interval. In the proposed two-timescale caching strategy, the long-term content placement is adaptive to the slowly-varying channel statistics, while the short-term beamforming and RRH clustering are adaptive to the instantaneous channel state information (CSI). The interrelationship between the single-timescale and two-timescale algorithms is discussed in detail.

3) Computer simulations are carried out to validate the effectiveness of the proposed algorithms. The necessity of the proposed single-timescale algorithms is also demonstrated, i.e., employing them as initialization methods for the two-timescale algorithm can accelerate its convergence remarkably. Moreover, we show that the proposed two-timescale algorithm outperforms those with heuristic caching strategies, such as the popularity-aware caching (PopC) and probabilistic caching (ProC) [22].

D. Organization of the Paper

The rest of the paper is organized as follows. In Section II, we present the system model of the content-aware C-RAN system and the corresponding problem formulation. In Section III, the proposed P-CCCP algorithm is developed along with its complexity analysis. In Section IV, we present the PDD-based algorithm and discuss its initialization. The two-timescale joint design algorithm is exposed in Section V. In Section VI, simulations are conducted to characterize the performance of the proposed algorithms. Finally, conclusions are drawn in Section VII.

Notations: Scalars, vectors and matrices are respectively denoted by lower case, boldface lower case and boldface upper case letters. For a matrix $\mathbf{A}$, $\mathbf{A}^T$ and $\mathbf{A}^H$ denote its transpose and conjugate transpose respectively, while $\mathbf{A} \succeq \mathbf{0}$ means that $\mathbf{A}$ is a positive semidefinite (square) matrix. The operators $\|\cdot\|$ and $\|\cdot\|_\infty$ denote the Euclidean and infinity norms of a complex vector, respectively. $\mathbb{C}^{m\times n}(\mathbb{R}^{m\times n})$ denotes the space of $m \times n$ complex (real) matrices. The set difference is defined as $\mathcal{A}\backslash\mathcal{B} \triangleq \{x|x \in \mathcal{A}, x \notin \mathcal{B}\}$.

II. System Model and Problem Formulation

In this section, we first describe the content-aware C-RAN system under study, in which downlink users are each served by a potentially overlapping cluster of RRHs. We then formulate the joint optimization problem for the content placement, RRH clustering and beamformer design.

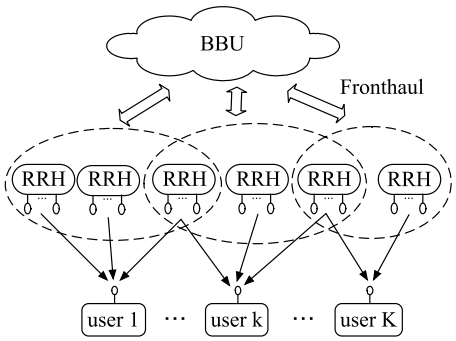


Fig. 1. System model of the content-aware C-RAN.

A. System Model

We consider a content-aware C-RAN, which consists of N multi-antenna RRHs, indexed by $n \in \mathcal{N} \triangleq \{1, \cdots, N\}$, K single-antenna mobile users, indexed by $k \in \mathcal{K} \triangleq \{1, \cdots, K\}$, and a centralized BBU pool, as shown in Fig. 1. The RRHs, each equipped with a common number L of antennas for simplicity, are individually connected to the BBUs via high-speed fronthaul links. We assume that the BBU pool has access to the information contents that can be potentially requested by all the users, and distributes each user's content to an individually selected cluster of RRHs via the fronthaul links. Each user is then cooperatively served by the associated RRH cluster through joint beamforming.

Let $\mathbf{w}_{k,n} \in \mathbb{C}^{L\times 1}$ denote the dowlink beamforming vector from RRH n to user k, and let $\mathbf{w}_k = [\mathbf{w}_{k,1}^H, \mathbf{w}_{k,2}^H, \cdots, \mathbf{w}_{k,N}^H]^H$ denote the aggregate, network wide beamforming vector form all RRHs to user k. In a similar way, let $\mathbf{h}_k^H$ denote the aggregate, network wide channel vector between the antennas of all the RRHs and user k. At a given symbol transmission instance, the received signal at user k can be written as

$$y_k = \mathbf{h}_k^H \mathbf{w}_k x_k + \sum_{j\in\mathcal{K}\backslash\{k\}} \mathbf{h}_k^H \mathbf{w}_j x_j + n_k, \tag{1}$$

where x_k (x_j) is the information symbol transmitted to user k ($j \neq k$) and n_k represents an additive white Gaussian noise term. Modeling $\{x_k\}$ and $\{n_k\}$ as zero-mean, mutually independent random variables and assuming that the channel vector $\mathbf{h}_k$ remains approximately constant over a transmission interval as defined below (i.e. block fading model), the signal-to-interference-plus-noise ratio (SINR) of user k can be defined as

$$\text{SINR}_k \triangleq \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum\limits_{j\in\mathcal{K}\backslash\{k\}} |\mathbf{h}_k^H \mathbf{w}_j|^2 + \sigma_k^2}. \tag{2}$$

Consequently, the achievable data rate of user k is given by $R_k = B\log_2(1+\text{SINR}_k)$, where B denotes the total available channel bandwidth.

Different from the conventional C-RAN framework, we here assume that each RRH can cache a certain amount of content

objects within a local storage device. At regular time intervals, referred to as transmission times, each user submits a content request according to a certain probability distribution specific to that user. If the requested content has already been cached locally at a serving RRH, then this RRH can access the content directly and transmit it to the user without the need for fronthaul data transfer.[4] It is assumed that enough time slots are available within a transmission time interval to complete the content delivery to the users, prior to the next transmission time. Content placement and delivery are controlled and managed by a cloud server, which consists of centralized BBUs, large data storage, software operated switches, etc.; the interested reader may consult [24], [33] for additional details about related implementation aspects. Without significant loss in generality, let us assume that the complete set of available user contents is represented by F binary files, indexed by $f \in \mathcal{F} = \{1, 2, \cdots, F\}$, each with normalized size of unity. The local storage size of RRH n is denoted as $Y_n \leq F$, which means that RRH n can cache Y_n content files at most. Let $c_{f,n} = 1$ indicate that content f is cached in RRH n and $c_{f,n} = 0$ otherwise, with the constraint that $\sum_{f\in\mathcal{F}} c_{f,n} \leq Y_n$.[5] Considering that a request for a content file that is not locally cached leads to a fronthaul utilization of one unit per serving RRH, the total fronthaul traffic reduction of the cache-enabled C-RAN can be expressed as [23]

$$C_B(s_{k,n}, c_{f,n}) = \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}} s_{k,n} \sum_{f\in\mathcal{F}} P_{k,f} c_{f,n}, \tag{3}$$

where $P_{k,f}$ denotes the probability that user k requests content file f and $s_{k,n}$ is the user-RRH association indicator, where $s_{k,n} = 1$ means that RRH n belongs to the serving cluster for user k and $s_{k,n} = 0$ otherwise. The probabilities $P_{k,f}$ can be viewed as the user preference indicators; here, it is assumed that they can be predicted or estimated via learning procedures which fall outside the scope of this work, see e.g., [34]–[36]. The cost of the transmission of the requested contents by all of the users from their serving RRHs can be assessed in terms of the total transmission power, defined here as

$$C_P(\mathbf{w}_{k,n}) = \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}} \|\mathbf{w}_{k,n}\|^2, \tag{4}$$

where it is assumed that $\mathbf{w}_{k,n} = \mathbf{0}$ when $s_{k,n} = 0$ (see also (6e)).

In this work, we introduce a new objective function, termed *caching efficiency* and defined as

$$C(s_{k,n}, c_{f,n}, \mathbf{w}_{k,n}) \triangleq \frac{C_B(s_{k,n}, c_{f,n})}{C_P(\mathbf{w}_{k,n})}, \tag{5}$$

which measures the amount of fronthaul traffic reduction that can be achieved per unit of consumed transmission power. The main motivation to employ the caching efficiency as the objective function in system design is based on the observation that with increasing transmission power budget, larger serving clusters can be formed for each user, which further reduces the fronthaul utilization. Hence, the objective function (5) is intuitively pleasing since it takes into account the proportionality relationship between the available power budget and the fronthaul reduction. Furthermore, if more emphasis on C_B or C_P is preferred, we can always put a proper weight (i.e., via the use of an exponent) on the denominator of (5). Note that generally, besides the above transmit beamforming power, the total power consumption also includes additional terms reflecting the constant power consumption of transceiver components induced by digital/analog signal processing, such as digital-to-analog converters (DAC), mixers, frequency synthesizers, etc. Since the considered caching efficiency is different from the conventional energy efficiency and due to the additional QoS constraints, these constant terms will not affect the algorithm design (which has been verified by our simulations); consequently, they are not considered in the sequel.

B. Problem Formulation

In this work, we aim to jointly optimize content placement, RRH clustering and cooperative beamforming at each transmission time interval, so as to maximize the caching efficiency, which can be formulated as the following optimization problem:

$$\max_{\{\mathbf{w}_{k,n}, s_{k,n}, c_{f,n}\}} \quad C(s_{k,n}, c_{f,n}, \mathbf{w}_{k,n}) \tag{6a}$$

$$\text{s.t. } \text{SINR}_k \geq \gamma_k, \ \forall k, \tag{6b}$$

$$s_{k,n}(1 - s_{k,n}) = 0, \quad \forall k, n, \tag{6c}$$

$$\sum_{k\in\mathcal{K}} s_{k,n} \leq X_n, \quad \forall n, \tag{6d}$$

$$(1 - s_{k,n})\mathbf{w}_{k,n} = \mathbf{0}, \quad \forall k, n, \tag{6e}$$

$$c_{f,n} = 0 \text{ or } 1, \quad \forall f, n, \sum_{f\in\mathcal{F}} c_{f,n} \leq Y_n, \quad \forall n. \tag{6f}$$

The QoS constraint (6b) requires that the SINR of user k should be no smaller than a given positive target threshold γ_k. Constraint (6c) means that the values of the user association indices $s_{k,n}$ can only be 0 or 1. Constraint (6d) indicates that the maximum number of users that RRH n can serve is limited by X_n. Finally, constraint (6e) forces the beamforming vector $\mathbf{w}_{k,n}$ to be an all-zero vector if user k is not served by RRH n.

In practice, content caching is typically updated over a larger timescale than RRH clustering and beamforming, in order to reduce the potential overhead brought up by content placement. Specifically, the long-term variables, i.e., the set of content placement indicators $\{c_{f,n}\}$, are adaptive to the slowly-varying statistics of the random system state $\{\mathbf{h}_k\}$. The short-term variables, i.e., the beamforming vectors $\{\mathbf{w}_k\}$ and RRH clustering indicator variables $\{s_{k,n}\}$, are adaptive to the realization of the system state, which changes at a faster rate. To account for this dual timescale, it is proposed to update $\{c_{f,n}\}$ only once per frame, which consists of N_s shorter time slots, and update $\{\mathbf{w}_k\}$ and $\{s_{k,n}\}$ in each

[4] In a C-RAN without caching capabilities, the RRHs need to fetch the requested content from the BBU pool via fronthaul links, and possibly from the cloud server via backhaul links. Since the backhaul/fronthaul links are usually implemented with dedicated fibers or free space optical links, the communication delays to fetch these contents can be safely ignored for simplicity.

[5] For simplicity, uncoded caching is considered in this work, while the investigation of coded caching and its impacts on RRH cooperation and beamforming remain an interesting avenue for future work.

time slot. Consequently, we also consider the following stochastic optimization problem:

$$\min_{\{\mathbf{w}_{k,n},s_{k,n},c_{f,n}\}} f(c_{f,n},\boldsymbol{\Theta}) \triangleq \mathbb{E}_{\{\mathbf{h}_k\}}(-C(s_{k,n},c_{f,n},\mathbf{w}_{k,n})) \quad \text{s.t. (6b)} - \text{(6f)}, \tag{7}$$

where $\boldsymbol{\Theta} \triangleq \{s_{k,n}, \mathbf{w}_{k,n}, \forall \mathbf{h}_k\}$ denotes the collection of all short-term variables for all possible system states.

Note that problems (6) and (7) are both highly nonconvex and nonlinear fractional programs, featuring both continuous and discrete variables which are coupled together in (6e) due to the RRH clustering operation. Furthermore, for problem (7), the long-term variables $\{c_{f,n}\}$ and short-term variables $\{s_{k,n}, \mathbf{w}_{k,n}\}$ are non-trivially coupled in the objective function. Consequently, problems (6) and (7) are quite challenging and it does not appear possible to obtain globally optimal solutions.[6] Meanwhile, we can readily see that the optimal solution of problem (6) at each time slot naturally provides an upper bound to that of problem (7) over multiple time slots. In the following, we first present two approaches to address problem (6), each one leading to a different numerical algorithm. Subsequently, we show that the proposed algorithms can be easily adapted to solve problem (7).

Remark 1: In the case that the requested files have different sizes, we can modify the total fronthaul traffic reduction C_B and the local storage size constraint as $C_B(s_{k,n},c_{f,n}) = \sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}} s_{k,n}\sum_{f\in\mathcal{F}} v_f P_{k,f} c_{f,n}$ and $\sum_{f\in\mathcal{F}} v_f c_{f,n} \leq Y_n$, respectively, where v_f denotes the size of file f. Since $\{v_f\}$ are known constants, the proposed algorithms are still applicable in this case. Besides, by fixing the Dinkelbach variable, it will be shown in the following that the proposed algorithms are also effective when the weighted sum of C_P and C_B is considered as the objective function, i.e., $C_P - \chi C_B$, where $\chi > 0$ is a predefined constant.

III. Proposed P-CCCP Algorithm

In this section, in order to make problem (6) tractable, we propose to first transform its fractional objective function into a numerator-denominator subtractive form by employing the Dinkelbach method [27]; we then present a P-CCCP algorithm to address the resulting subproblem. The proposed algorithm is motivated by the observation that by properly introducing auxiliary variables and penalizing certain constraints into the objective function, the pivotal coupling constraint (6e) can be expressed as a difference of convex (DC) function. Thus, we can use the CCCP method together with the block coordinate descent (BCD) method to iteratively solve the resultant subproblems. This leads to a twin-loop algorithm structure, where the inner loop seeks to approximately solve the penalized subproblem while the outer loop updates the penalty parameter and the Dinkelbach variable.

[6]Actually, these problems are NP-hard as shown for a simpler problem in [37, Theorem 1].

A. Reformulation of Problem (6)

By employing the Dinkelbach method, problem (6) can be reformulated as

$$\max_{\{\mathbf{w}_{k,n},s_{k,n},c_{f,n}\},\,\varsigma} C_B(s_{k,n},c_{f,n}) - \varsigma C_P(\mathbf{w}_{k,n}) \quad \text{s.t. (6b)} - \text{(6f)}, \tag{8}$$

where $\varsigma \in \mathbb{R}$ is the Dinkelbach variable. The main motivation behind this method is to convert the fractional objective in (6a) into a subtractive form that can be tackled more easily. It can be shown that there exists ς such that the optimal solution of (8) corresponds to that of (6). Specifically, let $P(\varsigma)$ denote the optimal value of the cost $C_B(s_{k,n},c_{f,n}) - \varsigma C_P(\mathbf{w}_{k,n})$ in (8) for a given ς. Then the maximum caching efficiency as per problem (6) is achieved for a $\varsigma = \varsigma^*$, where the latter satisfies $P(\varsigma^*) = 0$. To the best of our knowledge, existing algorithms available for related but simpler design problems in cache-enabled C-RAN, such as those in [22], [23], cannot be directly applied to solve the more general problem (8), even with fixed Dinkelbach variable. We next proceed with the further simplification of problem (8).

To begin, we note that (6c) represents a nonlinear equality constraint which is difficult to handle. To address this difficulty, we propose to relax this constraint by introducing auxiliary variables $\{\eta_{k,n}\}$ satisfying

$$s_{k,n}(1-s_{k,n}) \leq \eta_{k,n},\ \forall k,n. \tag{9}$$

In order to tighten this relaxation, the following constraint is also necessary:

$$s_{k,n}(1-s_{k,n}) \geq 0,\ \forall k,n. \tag{10}$$

In a similar way, the bilinear equality constraint (6e) can be relaxed by introducing auxiliary variables $\{t_{k,n}\}$ satisfying

$$\|\mathbf{w}_{k,n}\|^2 \leq t_{k,n},\ \forall k,n. \tag{11}$$

Accordingly, problem (8) can be transformed into the following problem:

$$\max_{\mathcal{W},\,\varsigma}\ C_B(s_{k,n},c_{f,n}) - \varsigma\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}} t_{k,n} - \beta\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\eta_{k,n} \tag{12a}$$

$$\text{s.t. (6b), (6d), (6f), (9), (10)}, \tag{12b}$$

$$\|\mathbf{w}_{k,n}\|^2 \leq s_{k,n}t_{k,n},\ \forall k,n, \tag{12c}$$

where β is a penalty parameter associated to the relaxed constraint (9) and $\mathcal{W} \triangleq \{\{\mathbf{w}_{k,n}\}, \{s_{k,n}\}, \{c_{f,n}\}, \{\eta_{k,n}\}, \{t_{k,n}\}\}$ denotes the extended set of search variables for notational simplicity. Note that the auxiliary variable $\eta_{k,n}$ in (9) can be viewed as a measure of the extent to which the original constraint (6c) is violated. Alternatively, the set of conditions $\{\eta_{k,n} = 0,\ \forall k,n\}$ can be viewed as a feasibility indicator, revealing that the RRH clustering constraints are satisfied. Hence, in light of (10), increasing the penalty parameter β forces the variables $\eta_{k,n}$ towards 0, which in the limit is equivalent to enforcing (6c). In the same way, the constraint (6e) is implicitly included in problem (12) through the combination of (12c) and the Dinkelbach variable.

Next, focusing on constraints (6b) and (12c), we find that problem (12) can be equivalently rewritten as

$$\max_{\mathcal{W},\,\varsigma}\ C_B(s_{k,n},c_{f,n})-\varsigma\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}t_{k,n}-\beta\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\eta_{k,n} \quad (13a)$$

$$\text{s.t.}\ \sqrt{\sum_{j\in\mathcal{K}\setminus\{k\}}|\mathbf{h}_k^H\mathbf{w}_j|^2+\sigma_k^2}\le\frac{1}{\sqrt{\gamma_k}}\mathbf{h}_k^H\mathbf{w}_k,\ \forall k, \quad (13b)$$

$$s_{k,n}\le\eta_{k,n}+s_{k,n}^2,\ \forall k,n, \quad (13c)$$

$$s_{k,n}-s_{k,n}^2\ge 0,\ \forall k,n, \quad (13d)$$

$$\text{(6d) and (6f)}, \quad (13e)$$

$$\sqrt{\|\mathbf{w}_{k,n}\|^2+\frac{1}{4}(s_{k,n}-t_{k,n})^2}\le\frac{1}{2}(s_{k,n}+t_{k,n}),\ \forall k,n.$$

Note that in (13b), $\mathbf{h}_k^H\mathbf{w}_k$ is implicitly restricted to the positive real domain, which incurs no loss of optimality since we can always phase-rotate the vector $\mathbf{w}_k$ such that $\mathbf{h}_k^H\mathbf{w}_k>0$ without affecting the cost function or the constraints. The transformation from (12c) to (13e) is based on the identity $\frac{1}{4}(s_{k,n}+t_{k,n})^2-\frac{1}{4}(s_{k,n}-t_{k,n})^2=s_{k,n}t_{k,n}$.

B. Algorithm Design

In the inner loop of the P-CCCP algorithm proposed here, we combine the BCD method and the CCCP to solve problem (13) with fixed Dinkelbach variable ς and penalty parameter β. It is worth mentioning that in order to generate a sequence of feasible solutions, a feasible initial solution is required by many iterative algorithms. In general, infeasible initial points contaminate the intermediate solutions obtained through the iterative optimization cycles, and often result in incorrect local optima. However, finding a "good" feasible initial point is not a simple task and often requires the same amount of computational resources as solving the original problem. Indeed, finding a feasible initial point of a non-convex problem, such as our caching efficiency maximization problem (6), is in general NP-hard. In our approach, the use of relaxation, i.e. inequality constraint (9), along with incorporation of the penalty parameter β in the cost function allow us to bypass the requirement of a non-trivial initialization. Specifically, we simply propose to randomly generate the values of $\{c_{f,n}\}$ required for initialization and then feasible $\{s_{k,n},\mathbf{w}_{k,n}\}$ can be gradually obtained as β increases during the iterative process.

To proceed, let us introduce a set of selection matrices $\mathbf{J}_n\in\{0,1\}^{L\times NL}$ defined as

$$\mathbf{J}_n=[\mathbf{0}_{L\times(n-1)L},\mathbf{I}_{L\times L},\mathbf{0}_{L\times(N-n)L}]. \quad (14)$$

With the help of (14), we can see that constraint (13e) is equivalent to $\|[\mathbf{J}_n\mathbf{w}_k,\frac{1}{2}(s_{k,n}-t_{k,n})]\|\le\frac{1}{2}(s_{k,n}+t_{k,n})$, which is indeed a second-order cone (SOC) constraint. Besides, we can approximate constraint (13c) in the ith inner iteration by replacing the concave part $-s_{k,n}^2$ by its first-order Taylor expansions around the current point $s_{k,n}^{(i)}$, which can be expressed as

$$s_{k,n}-\eta_{k,n}-\left(s_{k,n}^{(i)2}+2s_{k,n}^{(i)}\left(s_{k,n}-s_{k,n}^{(i)}\right)\right)\le 0. \quad (15)$$

Thus, in the ith inner iteration of the proposed P-CCCP algorithm, we solve the following problem:

$$\max_{\mathcal{W}}\ C_B(s_{k,n},c_{f,n})-\varsigma\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}t_{k,n}-\beta\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\eta_{k,n} \quad (16)$$
$$\text{s.t. (6d), (6f), (13b), (13d), (13e) and (15)},$$

which is a convex problem with respect to $\{c_{f,n}\}$ and $\{s_{k,n},\mathbf{w}_{k,n},\eta_{k,n},t_{k,n}\}$=$\mathcal{W}\setminus\{c_{f,n}\}$ respectively, but not jointly. Therefore, we propose to solve problem (16) by a BCD-type method in which the blocks of optimization variables $\mathcal{W}\setminus\{c_{f,n}\}$ and $\{c_{f,n}\}$ are successively updated while keeping the other block fixed.

Step 1: The optimization of $\mathcal{W}\setminus\{c_{f,n}\}$ is a convex second-order cone program (SOCP), which can be expressed as

$$\max_{\mathcal{W}\setminus\{c_{f,n}\}}\ C_B(s_{k,n},c_{f,n})-\varsigma\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}t_{k,n}-\beta\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\eta_{k,n}$$
$$\text{s.t. (6d), (13b), (13d), (13e) and (15)}. \quad (17)$$

This problem can be efficiently handled by off-the-shelf solvers, e.g. [38].

Step 2: With fixed $\mathcal{W}\setminus\{c_{f,n}\}$, we have the following subproblem (separable among different n):

$$\max_{\{c_{f,n}\}}\ \sum_{f\in\mathcal{F}}\kappa_{f,n}c_{f,n}$$
$$\text{s.t. } c_{f,n}=0\text{ or }1,\ \forall f,\ \sum_{f\in\mathcal{F}}c_{f,n}\le Y_n, \quad (18)$$

where $\kappa_{f,n}=\sum_{k\in\mathcal{K}}s_{k,n}P_{k,f}$ is the amount of fronthaul reduction if RRH n caches file f, i.e., the benefit of caching file f at RRH n. In essence, the aim of problem (18) is to determine which subset of Y_n files should be cached by RRH n. The optimal solution to such a problem is simply to cache the Y_n files that have the largest benefits, i.e.,

$$c_{f,n}^{\text{opt}}=\begin{cases}1, & \text{if } f\in\mathcal{K}_n\\ 0, & \text{otherwise}\end{cases} \quad (19)$$

where $\mathcal{K}_n\triangleq\underset{\bar{\mathcal{K}}\subset\mathcal{F},|\bar{\mathcal{K}}|=Y_n}{\arg\max}\left(\sum_{f\in\bar{\mathcal{K}}}\kappa_{f,n}\right)$.

In the outer loop, the Dinkelbach variable can be updated as

$$\varsigma=(C_B(s_{k,n},c_{f,n})/C_P(\mathbf{w}_{k,n}))^i, \quad (20)$$

where i denotes the inner iteration index, while the penalty parameter β can be updated according to $\beta=\min\{\mu\beta,\beta^{\max}\}$, where $\mu>1$ is a control parameter that increases the penalty by a fixed proportion during each outer iteration and $\beta^{\max}$ denotes the maximum penalty value.

The proposed P-CCCP algorithm to solve problem (6) is summarized as Algorithm 1. It is worth noting that problem (16) with fixed $\{c_{f,n}\}$ involves $K(N+1)$ linear constraints and $K(2N+1)$ SOC constraints, which consist of K SOCs of dimension $K+1$, KN SOCs of dimension 2 and KN SOCs of dimension 3. The number of optimization variables is on the order of $\kappa=\mathcal{O}(KNL+3KN)$. By applying basic elements of complexity analysis as in [39], the complexity of a generic interior-point method (IPM) for solving problem

Algorithm 1 The Proposed P-CCCP Algorithm

Initialize $\{c_{f,n}\}^0$ β_0, $\beta^{\max}$ and ς_0. Set $m=0$.
repeat
  Set the inner iteration number $i=0$.
  repeat
    Solve problem (17) with fixed $\{c_{f,n}\}^i$ to obtain the updated $\{\mathcal{W}\backslash\{c_{f,n}\}^i\}^{i+1}$.
    Obtain $\{c_{f,n}\}^{i+1}$ from (18).
    $i \leftarrow i+1$.
  until some convergence condition is met.
  Set $\beta=\min\{\mu\beta,\beta^{\max}\}$ and update ς according to (20).
  Assign $\mathcal{W}^i$ to $\mathcal{W}^0$ and set $m \leftarrow m+1$.
until some convergence condition is met.

(16) can be expressed as $\mathcal{O}(\kappa\sqrt{5KN+N+2K}(KN+N+\kappa(KN+N)+K(K+1)^2+13KN+\kappa^2))$. Therefore, by letting $K=N=L\rightarrow\infty$, the worst-case asymptotic complexity of Algorithm 1 can be evaluated as $\mathcal{O}(MIN^{10})$, where M and I denote the maximum number of outer and inner iterations, respectively.

Remark 2: The proposed Algorithm 1 combines the benefits of the Dinkelbach, the BCD and the CCCP methods. By introducing the Dinkelbach variable ς, the caching efficiency maximization problem (6) is transformed into a series of Dinkelbach subproblems. Then, the BCD and CCCP method are employed to solve the latter subproblems in the inner loop; subsequently, the refined variables, i.e. $\{s_{k,n},c_{f,n},\mathbf{w}_{k,n}\}$ are used to update the Dinkelbach variable ς in the outer loop.

Remark 3: If one wants to put more priority on C_B or C_P, the following objective function can be considered: $C(s_{k,n},c_{f,n},\mathbf{w}_{k,n})=\frac{C_B(s_{k,n},c_{f,n})}{\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\|\mathbf{w}_{k,n}\|^{2p}}$, where p is a predefined constant, so that $0<p<1$ puts more emphasis on C_B while $p>1$ has the opposite effect. Correspondingly, the proposed Algorithm 1 can be easily modified to address the resultant problem. To be specific, we only need to change the terms involving $t_{k,n}$ in the objective of problem (16) from $-\varsigma\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}t_{k,n}$ to $-\varsigma\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}t_{k,n}^p$. Then, if $p>1$, the resulting problem (17) is still a convex problem, which is easy to address. If $0<p<1$, we can see that $t_{k,n}^p$ is concave with respect to $t_{k,n}$ (because $t_{k,n}$ is generally non-negative), therefore, one can employ the CCCP concept to approximate it with a convex surrogate function. The details are omitted here for brevity.

Remark 4: A complete characterization of the convergence properties of Algorithm 1 is rather involved and falls outside the scope of this work, where the focus is on algorithm design and performance study. Therefore, the detailed proof is left for future research.

IV. Proposed PDD-Based Algorithm

In the previous section, we proposed the P-CCCP algorithm which, at each iteration, requires the solution of an SOCP problem. In general, this type of approach is characterized by relatively high computational complexity since off-the-shelf software solvers must be employed. From another perspective, this suggests that the problem structure is not fully exploited in the proposed P-CCCP algorithm. In this section, motivated by these considerations, we develop an alternative PDD-based algorithm that is more efficient and simpler to implement. The proposed PDD-based algorithm also relies on the Dinkelbach method and exhibits a twin-loop structure: in this case however, the inner loop seeks to (approximately) solve an augmented Lagrangian (AL) problem (see e.g., [40]–[42]) using a block minimization technique, while the outer loop updates the Dinkelbach variable and either the dual variables or the penalty parameter, depending on a constraint violation status. Interestingly, we show that each subproblem in the inner loop's block minimization can be solved either in closed-form or by the bisection method [43].

A. Reformulation of Problem (8)

We first rewrite problem (8) as follows:

$$\min_{\{\mathbf{w}_{k,n},s_{k,n},c_{f,n}\},\,\varsigma} -C_B(s_{k,n},c_{f,n})+\varsigma C_P(\mathbf{w}_{k,n}) \quad \text{s.t. (6b)} - \text{(6f)}. \tag{21}$$

Next, we introduce auxiliary variables $\{\hat{s}_{k,n}\}$ and $\{\mathbf{w}_k^j\}_{j\in\mathcal{K}\backslash\{k\}}$ which satisfy

$$\mathbf{w}_k^j=\mathbf{w}_k,\ \forall j\in\mathcal{K}\backslash\{k\},\ \forall k, \tag{22}$$

$$s_{k,n}=\hat{s}_{k,n},\ \forall k,n. \tag{23}$$

Note that (22) and (23) can be interpreted as introducing $K-1$ and 1 redundant copies of variables $\mathbf{w}_k$ and $s_{k,n}$, respectively.[7] Then, problem (21) can be equivalently expressed as

$$\min_{\bar{\mathcal{W}},\,\varsigma} -C_B(s_{k,n},c_{f,n})+\varsigma C_P(\mathbf{w}_{k,n}) \tag{24a}$$

$$\text{s.t. (6d)} - \text{(6f)},\ \text{(22) and (23)}, \tag{24b}$$

$$\frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{\sum_{j\in\mathcal{K}\backslash\{k\}}|\mathbf{h}_k^H\mathbf{w}_j^k|^2+\sigma_k^2}\geq\gamma_k,\ \forall k, \tag{24c}$$

$$s_{k,n}(1-\hat{s}_{k,n})=0,\ \forall k,n, \tag{24d}$$

$$0\leq\hat{s}_{k,n}\leq 1,\ \forall k,n, \tag{24e}$$

where $\bar{\mathcal{W}}\triangleq\{\{\mathbf{w}_k\},\{\mathbf{w}_k^j\}_{j\in\mathcal{K}\backslash\{k\}},\{s_{k,n}\},\{\hat{s}_{k,n}\},\{c_{f,n}\}\}$. The introduction of these auxiliary variables represents a critical step in developing the proposed PDD-based algorithm. Indeed, by adopting these new variables, we can partition the complete set of optimization variables into smaller non-overlapping subsets, or blocks, that can be optimized separately. Specifically, the joint optimization problem (21) can be decomposed into a number of subproblems which either admit closed-form solutions or can be solved via simple iterative approaches. Hence, through the introduction of auxiliary variables and judicious use of the block structure, low-complexity algorithms can be devised for the optimization of each block of variables, so that ultimately, the underlying problem (6) can be easily solved.

[7]We emphasize that in contrast to $s_{k,n}$ which only takes on binary values, its copy $\hat{s}_{k,n}$ is a continuous variable.

B. Algorithm Design

In this subsection, we aim to conceive an efficient PDD-based algorithm to solve problem (24). To this end, the AL of problem (24) is first formulated as

$$\begin{aligned}\min_{\bar{\mathcal{W}},\,\varsigma}\ & -C_B(s_{k,n},c_{f,n})+\varsigma\sum_{k\in\mathcal{K}}\|\mathbf{w}_k\|^2+P_\rho\\ \text{s.t. }& \text{(6d), (6f), (24c) and (24e),}\end{aligned}\tag{25}$$

where the penalty term

$$\begin{aligned}P_\rho \triangleq\ & \frac{1}{2\rho}\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\Big((s_{k,n}(1-\hat{s}_{k,n})+\rho_m\lambda_{k,n})^2\\ &+(s_{k,n}-\hat{s}_{k,n}+\rho_m\hat{\lambda}_{k,n})^2\Big)\\ &+\frac{1}{2\rho}\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{K}\setminus\{k\}}\|\mathbf{w}_k-\mathbf{w}_k^j+\rho_m\boldsymbol{\mu}_{j,k}\|^2\\ &+\frac{1}{2\rho}\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k+\rho_m\boldsymbol{\xi}_{k,n}\|^2,\end{aligned}\tag{26}$$

and $\{\lambda_{k,n}\}$, $\{\hat{\lambda}_{k,n}\}$, $\{\boldsymbol{\mu}_{j,k}\}$ and $\{\boldsymbol{\xi}_{k,n}\}$ denote the dual variables corresponding to the constraints (24d), (23), (22) and (6e), respectively. The coefficient $\rho>0$ is used to control the size of the penalty (i.e., decreasing ρ increases the penalty).

Our proposed algorithm consists of two embedded loops. In the outer loop, indexed by positive integer m, we update the Dinkelbach variable ς_m and either the dual variables $\{\lambda_{k,n}^m,\hat{\lambda}_{k,n}^m,\boldsymbol{\mu}_{j,k}^m,\boldsymbol{\xi}_{k,n}^m\}$ or the penalty parameter ρ_m, where the dependence of these variables on iteration index m is now made explicit for clarity. In the inner loop, we employ the block successive upper-bound minimization (BSUM) method [44] to iteratively optimize the primal variables $\bar{\mathcal{W}}$ over selected blocks of variables while keeping the other variables fixed. In the following, we first develop the BSUM method in details, then present the update of the dual variables, Dinkelbach variable and penalty parameter, and finally summarize the overall PDD-based algorithm.

In the inner loop, with fixed values of $\{\lambda_{k,n}^m,\hat{\lambda}_{k,n}^m,\boldsymbol{\mu}_{j,k}^m,\boldsymbol{\xi}_{k,n}^m\}$, ρ_m and ς_m, we propose to divide the primal variables into four blocks that will be treated separately, i.e., $\{s_{k,n}\}$, $\{\hat{s}_{k,n}\}$, $\{\mathbf{w}_k,\mathbf{w}_k^j\}$ and $\{c_{f,n}\}$. We now proceed with the optimization of each block.

1. *Block* $\{s_{k,n}\}$: The optimization problem of $\{s_{k,n}\}$ can be expressed as

$$\begin{aligned}\min_{\{s_{k,n}\}}\ & -\sum_{k\in\mathcal{K}}s_{k,n}\sum_{f\in\mathcal{F}}P_{k,f}c_{f,n}\\ &+\frac{1}{2\rho_m}\sum_{k\in\mathcal{K}}\|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k+\rho_m\boldsymbol{\xi}_{k,n}^m\|^2\\ &+\frac{1}{2\rho_m}\sum_{k\in\mathcal{K}}\Big((s_{k,n}(1-\hat{s}_{k,n})+\rho_m\lambda_{k,n}^m)^2\\ &+(s_{k,n}-\hat{s}_{k,n}+\rho_m\hat{\lambda}_{k,n}^m)^2\Big)\\ \text{s.t. }&\sum_{k\in\mathcal{K}}s_{k,n}\le X_n,\ \forall n.\end{aligned}\tag{27}$$

It can be seen that for each n, the variables $\{s_{k,n}\}_{k\in\mathcal{K}}$ can be optimized separately in a parallel manner. In particular, a closed-form solution can be obtained for the optimal $\{s_{k,n}\}$, as explained in further details in Appendix A.

2. *Block* $\{\mathbf{w}_k,\mathbf{w}_j^k\}_{j\in\mathcal{K}\setminus\{k\}}$: The corresponding optimization problem can be expressed as[8]

$$\begin{aligned}\min_{\mathbf{w}_k,\{\mathbf{w}_j^k\}_{j\neq k}}\ & \varsigma_m\|\mathbf{w}_k\|^2\\ &+\frac{1}{2\rho_m}\sum_{n\in\mathcal{N}}\|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k+\rho_m\boldsymbol{\xi}_{k,n}^m\|^2\\ &+\frac{1}{2\rho_m}\sum_{j\in\mathcal{K}\setminus\{k\}}\Big(\|\mathbf{w}_k-\mathbf{w}_k^j+\rho_m\boldsymbol{\mu}_{j,k}^m\|^2\\ &+\|\mathbf{w}_j^j-\mathbf{w}_j^k+\rho_m\boldsymbol{\mu}_{k,j}^m\|^2\Big)\\ \text{s.t. }&\frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{\sum_{j\in\mathcal{K}\setminus\{k\}}|\mathbf{h}_k^H\mathbf{w}_j^k|^2+\sigma_k^2}\ge\gamma_k.\end{aligned}\tag{28}$$

Problem (28) can be efficiently solved by resorting to the Lagrangian dual problem and employing the bisection method. The detailed derivation is provided in Appendix B.

3. *Block* $\{\hat{s}_{k,n}\}$: We consider the following problem:

$$\begin{aligned}\min_{\hat{s}_{k,n}}\ & \tfrac{1}{2\rho_m}(s_{k,n}(1-\hat{s}_{k,n})+\rho_m\lambda_{k,n}^m)^2\\ &+\tfrac{1}{2\rho_m}(s_{k,n}-\hat{s}_{k,n}+\rho_m\hat{\lambda}_{k,n}^m)^2,\\ \text{s.t. }& 0\le\hat{s}_{k,n}\le 1.\end{aligned}\tag{29}$$

Problem (29) also admits a closed-form solution, as detailed in Appendix C.

4. *Block* $\{c_{f,n}\}$: The optimization of $\{c_{f,n}\}$ can be formulated as problem (18), the solution of which has already been described in Section III-B.

In the outer loop, the dual variables $\{\lambda_{k,n}^m,\ \hat{\lambda}_{k,n}^m,\ \boldsymbol{\mu}_{j,k}^m,\ \boldsymbol{\xi}_{k,n}^m\}$ can be updated by means of

$$\begin{aligned}\lambda_{k,n}^{m+1}&=\lambda_{k,n}^m+\frac{1}{\rho_m}(s_{k,n}(1-\hat{s}_{k,n})),\\ \hat{\lambda}_{k,n}^{m+1}&=\hat{\lambda}_{k,n}^m+\frac{1}{\rho_m}(s_{k,n}-\hat{s}_{k,n}),\\ \boldsymbol{\mu}_{j,k}^{m+1}&=\boldsymbol{\mu}_{j,k}^m+\frac{1}{\rho_m}(\mathbf{w}_k-\mathbf{w}_k^j),\\ \boldsymbol{\xi}_{k,n}^{m+1}&=\boldsymbol{\xi}_{k,n}^m+\frac{1}{\rho_m}((1-s_{k,n})\mathbf{J}_n\mathbf{w}_k).\end{aligned}\tag{30}$$

As for the Dinkelbach variable and the penalty parameter, they can be updated as

$$\varsigma_{m+1}=C_B(s_{k,n},c_{f,n})^m/C_P(\mathbf{w}_{k,n})^m,\ \rho_{m+1}=q\rho_m,\tag{31}$$

where $q<1$ is a control parameter used to increase the value of the penalty term P_ρ in (26) during each outer iteration. Besides, we denote the maximum constraint violation among all the equality constraints in problem (24) as ϖ, which is shown as follows:

$$\begin{aligned}\varpi=\max\nolimits_{\forall k,j,n}\ &\big\{|s_{k,n}(1-\hat{s}_{k,n})|,\ |s_{k,n}-\hat{s}_{k,n}|,\\ &\|\mathbf{w}_k-\mathbf{w}_k^j\|_\infty,\ \|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k\|_\infty\big\}.\end{aligned}\tag{32}$$

This important quantity can be employed to determine if the proposed algorithm converges, and whether we should update the dual variables or increase the penalty parameter.

The main steps of the proposed PDD-based algorithm are summarized in Algorithm 2. Similar to the complexity analysis

[8]Due to the additive nature of the AL, we only need to consider a single value of k at a time, i.e., optimization for other values of k can be done separately in parallel.

in Section III-B, we observe that the complexity for solving problems (18), (27) and (29) is almost negligible compared with that of solving problem (28). Overall, the complexity can be expressed as[9] $\mathcal{O}\left(MIN^3L^3K^4\log_2\left(\frac{\lambda_{\max}-\lambda_{\min}}{\varepsilon}\right)\right)$, where $\lambda_{\max}=\max\{\overline{\lambda}_k\}_{k\in\mathcal{K}}$ and $\lambda_{\min}=\min\{\underline{\lambda}_k\}_{k\in\mathcal{K}}$ denote the upper and lower bounds of the corresponding dual variables (see Appendix B) and ε denotes the precision of the bisection method. For completeness, a simple initialization method based on zero-forcing (ZF) beamforming [45] is proposed in Appendix D.

Remark 5: Similar to Algorithm 1 in Section III, we leave the detailed convergence proof of the proposed PDD-based algorithm for future research. If an exponential parameter is introduced in the objective in (28), i.e. if $\|\mathbf{w}_k\|^2$ is replaced by $\|\mathbf{w}_k\|^{2p}$, one can utilize a similar method as that in Remark 3 to modify Algorithm 2, i.e., employing the BSUM method to make subproblem (28) tractable.

Algorithm 2 The Proposed PDD-Based Algorithm

1: Initialize $\{\mathbf{w}_k^j\}^0, \{c_{f,n}\}^0, \{s_{k,n}\}^0=\{\hat{s}_{k,n}\}^0, \eta_0, \varrho_0, \rho_0, q$ and ς_0.
2: Set the outer iteration number $m=0$.
3: **repeat**
4: Set the inner iteration index $i=0$.
5: **repeat**
6: Update $\{\mathbf{w}_k, \mathbf{w}_j^k\}_{j\neq k}$ by solving problem (28) (Appendix B).
7: Update $\{\hat{s}_{k,n}\}$ by solving problem (29) (Appendix C).
8: Update $\{c_{f,n}\}$ by (19).
9: Update $\{s_{k,n}\}$ by solving problem (27) (Appendix A). $i\leftarrow i+1$.
10: **until** some convergence condition is met.
11: Assign $\bar{\mathcal{W}}^i$ to $\bar{\mathcal{W}}^0$. Calculate ϖ via (32). If $\varpi\leq\eta_m$ and update the dual variables via (30), otherwise set $\rho_{m+1}=q\rho_m$. Set $\varrho_{m+1}=q\varrho_m$, $\eta_{m+1}=\varrho_{m+1}^{1/6}$ and update the Dinkelbach parameter via (20).[10] $m\leftarrow m+1$.
12: **until** some convergence condition is met.

V. Two-Timescale Joint Design Algorithm

In this section, we propose a novel two-timescale joint design algorithm based on the TOSCA framework [32] to address problem (7). In the proposed two-timescale design, we update the long-term variables at the end of each frame and optimize the short-term variables based on the instantaneous CSI at each time slot, as shown in Fig. 2. Since the long-term variables $\{c_{f,n}\}$ are discrete and difficult to handle, we propose to relax them into continuous variables and introduce

[9]The main factor affecting the computational complexity of Algorithm 2 is the need to perform the eigenvalue decomposition of a $KNL\times KNL$ matrix multiple times. For general matrices, the associated complexity would be $\mathcal{O}(K^3N^3L^3)$ for each such eigendecomposition. However, since $\mathbf{A}_k$ and $\mathbf{D}_k$ (defined in Appendix B) are sparse matrices, the corresponding complexity can be significantly reduced by further exploiting the special structure of $\mathbf{A}_k$ and $\mathbf{D}_k$, which we shall not further detail in this work due to space limitation.

[10]Note that ϱ_m is a parameter that controls the decaying of the constraint violation, i.e., if $\varpi\geq\eta_m$ (the value of η_m is controlled by ϱ_m), then the penalty parameter ρ_{m+1} is further decreased to decrease ϖ.

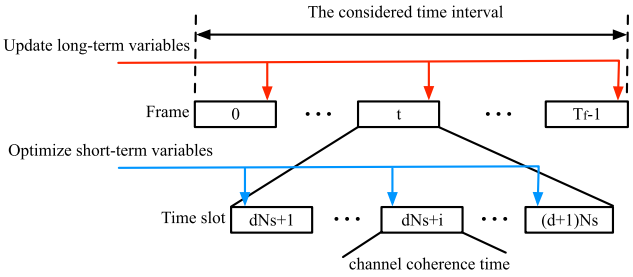


Fig. 2. Timeline (frame structure) of the considered two-timescale design.

a continuous smooth function into the objective to promote sparsity [46]. Specifically, we transform (7) into the following problem:

$$\min_{\{\mathbf{w}_{k,n},s_{k,n},c_{f,n}\}} \hat{f}(c_{f,n},\boldsymbol{\Theta})\triangleq f(c_{f,n},\boldsymbol{\Theta})-\sum_{f\in\mathcal{F},n\in\mathcal{N}} g_c(c_{f,n})$$
$$\text{s.t. (6b)}-\text{(6e)},\ c_{f,n}\in[0,1],\ \forall f,n,\ \sum_{f\in\mathcal{F}} c_{f,n}\leq Y_n, \tag{33}$$

where $g_c(\cdot)$ can be a smooth concave function, such as the logarithm or exponential.

It can be observed that problem (33) can be decomposed into a long-term master problem and a family of short-term subproblems, which can be expressed as

$$\mathcal{P}_L:\ \min_{\mathbf{c}}\ \hat{f}(\mathbf{c},\boldsymbol{\Theta}^*(\mathbf{c}))$$
$$\text{s.t. } c_{f,n}\in[0,1],\ \forall f,n,\ \sum_{f\in\mathcal{F}} c_{f,n}\leq Y_n,\ \forall n,$$
$$\mathcal{P}_S(\mathbf{c},\mathbf{h}_k):\ \min_{\{\mathbf{w}_{k,n}\},\,\mathbf{s}} -C(\mathbf{s},\mathbf{c},\mathbf{w}_{k,n})$$
$$\text{s.t. (6b)}-\text{(6e)},$$

where we have $\mathbf{c}\triangleq\{c_{f,n}\}$, $\mathbf{s}\triangleq\{s_{f,n}\}$, $\boldsymbol{\Theta}^*(\mathbf{c})=\{\mathbf{s}^*(\{\mathbf{h}_k\}),\mathbf{w}_{k,n}^*(\{\mathbf{h}_k\}),\forall\mathbf{h}_k\in\Omega\}$, $\{\mathbf{s}^*,\mathbf{w}_{k,n}^*\}$ denotes the solution of $\mathcal{P}_S(\mathbf{c},\mathbf{h}_k)$ and Ω denotes the sample space. As can be seen, the short-term problems $\mathcal{P}_S(\mathbf{c},\mathbf{h}_k)$ in each time slot can be solved by employing Algorithm 1 and Algorithm 2, where we just fix the variables in $\mathbf{c}$ and leave the other steps in these two algorithms unchanged. For the long-term problem $\mathcal{P}_L$, we introduce the following surrogate function for its objective:

$$\bar{f}(\mathbf{c})=\sum_{n\in\mathcal{N}}\sum_{f\in\mathcal{F}} q_{f,n}^t c_{f,n}$$
$$-\sum_{f\in\mathcal{F},n\in\mathcal{N}} g_c(c_{f,n})+\tau\|\mathbf{c}-\mathbf{c}^t\|^2, \tag{34}$$

where the superscript t denotes the frame index, $\mathbf{c}^t$ denotes the long-term variables obtained in the t-th frame, τ is a small positive number. $q_{f,n}^t$ can be recursively computed as

$$q_{f,n}^t=(1-\rho^t)q_{f,n}^{t-1}-\rho^t\sum_{i=tN_s+1}^{(t+1)N_s}\frac{\kappa_{f,n}^t(i)/C_P(\mathbf{w}_{k,n}^t(i))}{N_s}. \tag{35}$$

where i denotes the time slot index, $\kappa_{f,n}^t(i)$ represents the benefit of caching file f at RRH n in the i-th time slot of frame t, ρ^t is a sequence satisfying $\rho^t\to 0$, $1/\rho^t\leq\mathcal{O}(t^\beta)$, for some $\beta\in(0,1)$, and $\sum_t(\rho^t)^2<\infty$. Note that the statistical

Algorithm 3 The Proposed Two-Timescale Joint Design Algorithm

1: Initialize $\{\mathbf{w}_k^j\}^0, \{c_{f,n}\}^0, \{s_{k,n}\}^0, \rho^0, t = 0$.
2: **Step 1**: At each time slot i, solve problem $\mathcal{P}_S(\mathbf{c}, \mathbf{h}_k)$ using Algorithm 1 or Algorithm 2 and obtain solution $\{s_{k,n}(i), \mathbf{w}_{k,n}(i)\}$. Update the surrogate function $\bar{f}(\mathbf{c})$ using $\{s_{k,n}(i), \mathbf{w}_{k,n}(i)\}$.
3: **Step 2**: At the end of each frame, solve problem (36) to obtain $\bar{\mathbf{c}}^t$, update $\mathbf{c}^{t+1}$ by $\mathbf{c}^{t+1} = (1-\gamma^t)\mathbf{c}^t + \gamma^t \bar{\mathbf{c}}^t$, where $\gamma^t \in (0, 1]$ is a sequence satisfying $\sum_t \gamma^t = \infty$, $\sum_t (\gamma^t)^2 < \infty$.
4: Let $t = t + 1$ and return to Step 1.

information of the CSI is implicitly contained in $q_{f,n}^t$ through $\kappa_{f,n}^t(i)$ and $\mathbf{w}_{k,n}^t(i)$. At the end of each frame, we update the long-term variables by solving the following problem:

$$\min_{\mathbf{c}} \bar{f}(\mathbf{c}) \quad \text{s.t. } c_{f,n} \in [0,1],\ \forall f, n,\ \sum_{f \in \mathcal{F}} c_{f,n} \le Y_n,\ \forall n, \tag{36}$$

which is convex and can be easily solved. The proposed two-timescale joint design algorithm is summarized in Algorithm 3.

VI. Simulation Results

In this section, the performance of the proposed algorithms is evaluated numerically. The following system parameter values are used throughout unless otherwise specified: $N = 7$, $K = 12$, $L = 2$, $F = 1120$ and $\sigma_k^2 = -90$ dBm, $\forall k$.[11] Each RRH is located at the center of a hexagonal-type cell, where the propagation distance between adjacent RRHs is set to 100 meters and the users are uniformly and independently distributed in the area. We consider Rayleigh fading channels with large-scale pathloss (in dB) modeled as $-147.3 - 43.3\log_{10} D$, where the distance D is measured in kilometers. For simplicity, we assume that all users have the same SINR requirements, and that each RRH has the same local storage size and can support the same number of users, i.e.: $\gamma_k = \gamma, \forall k \in \mathcal{K}$, $X_n = X$, $Y_n = Y$, $\forall n \in \mathcal{N}$. For comparison, we also provide the results of two separate design algorithms, namely the McCormick envelopes Branch-and-Bound (ME-BB) algorithm [49] and the separate PDD-based algorithm, where the discrete variables $\{s_{k,n}, c_{f,n}\}$ and the beamforming vectors $\{\mathbf{w}_k\}$ are separately optimized. In Algorithm 1, we use the following parameter values: $\mu = 0.85$, $\beta^{\max} = 100$, $\beta_0 = 0.1$ and $\varsigma_0 = 1$, while in Algorithm 2, we set $\rho_0 = 20$, $\eta_0 = 100$, $\varrho_0 = 100$, $q = 0.95$ and $\varsigma_0 = 1$. The convex problem (16) in Algorithm 1 is solved by CVX [38] and the integer programming problem in the ME-BB algorithm is solved by the MOSEK solver [50]. The simulations are carried out on a computer with Intel (i7-6700HQ) CPU running at 2.60GHz and with 8GB RAM.

[11] More users can be supported by performing proper user scheduling, e.g. according to the traffic demand and channel quality, etc. In this work, for simplicity, we assume that the proposed algorithms operate under the condition of fixed user scheduling, as in [22], [47], [48]. Further investigation into user scheduling and the influence of inactive users on content placement are left for future work.

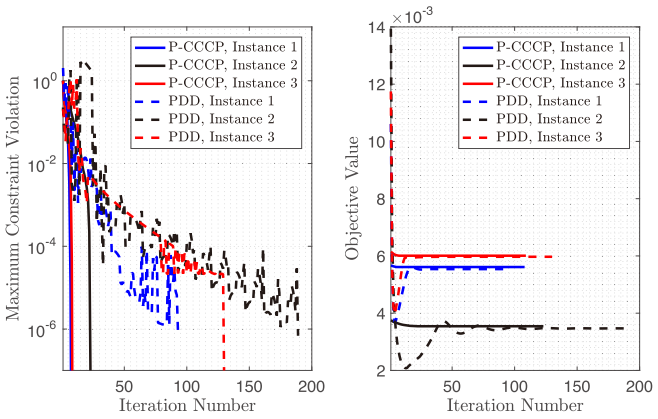


Fig. 3. Maximum constraint violation and objective value versus outer iteration number for 3 realizations.

In the simulations, we assume that the popularity of the files can be measured based on the number and behavior of the requests. According to [22], [23], [51], in practice, the mobile user requests usually follow a certain content popularity distribution, e.g., the Zipf distribution. Therefore, in our simulations, we assume that four types of files can be requested, all with similar Zipf distribution with parameter 0.4. Besides, four types of users with different file preferences are considered, i.e.: Type l ($l \in \{1, 2, 3, 4\}$) users prefer Type l files with probability 0.4 and the other three types of files with probability 0.2.

In Fig. 3, we illustrate the convergence behavior of the proposed Algorithm 1 (P-CCCP) and Algorithm 2 (PDD) for three different realizations in the case of $X = 6$, $Y = 200$ and $\gamma = 6$dB. The plots show that both algorithms can reach convergence within a few hundred outer iterations. From these and other instances, it is observed that Algorithm 1 tends to converge slightly faster than Algorithm 2, although their steady state performance is comparable on average. This is apparently due to the fact that the latter uses an off-the-shelf solver to simultaneously optimize the search variables in (17), whereas the former relies on block minimization, in which the search variables are divided into several smaller blocks.

Next, in Fig. 4 and 5, we illustrate the effect of content-aware RRH clustering in the case of $K = 12$, $F = 350$, $X = 6$ and $Y = 50$. Here, for more clarity, we consider two types of files and users with different file preferences, i.e.: Type l ($l \in \{1, 2\}$) users prefer Type l files with probability 0.8 and the other type of files with probability 0.2. In Fig. 4 (a) and (b), the RRH clustering is jointly optimized along with content placement using the proposed Algorithms 1 and Algorithm 2. It can be observed that in order to maximize caching efficiency as defined in (5), the RRHs selected to serve a particular user are not necessarily the closest one to that user (i.e. a distant RRH may be selected while a nearby one is not). It can also be observed that the final RRH clustering obtained with Algorithms (1) and Algorithm 2 are quite similar. The differences can be apparently explained on the basis of the differences between these two algorithms, especially that they can produce different stationary solutions with similar performance (in terms of objective value). In Fig. 4 (c) and (d), we present the

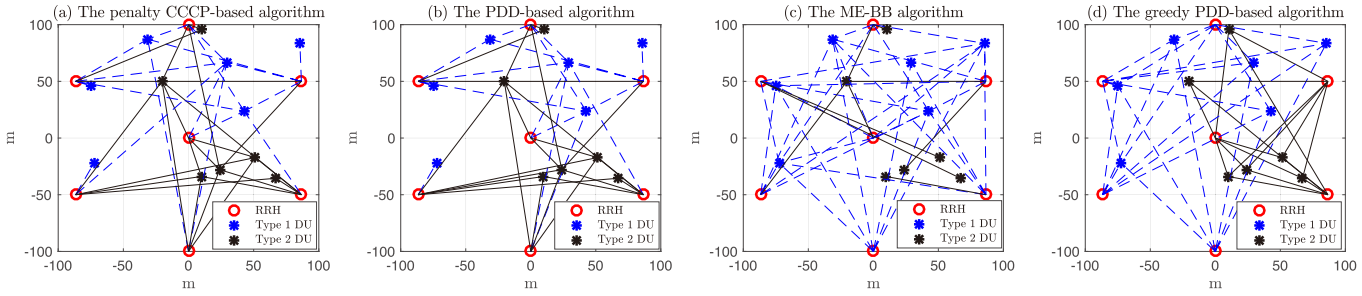


Fig. 4. RRH clustering with caching obtained by algorithm 1, Algorithm 2, the ME-BB and separate PDD-based algorithms.

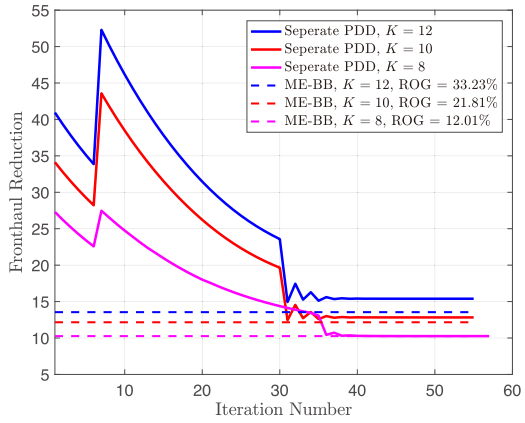


Fig. 5. Fronthaul reduction performance comparison between the ME-BB algorithm and the separate PDD-based algorithm.

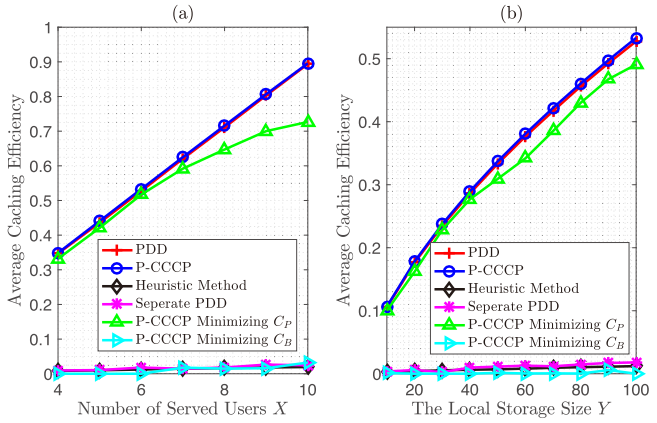


Fig. 6. Average caching efficiency versus the maximum number of concurrent users X served by each RRH (with fixed $Y = 100$) and the local storage size Y of each RRH (fixed $X = 6$).

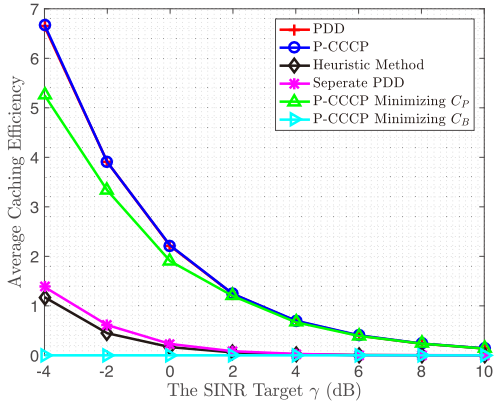


Fig. 7. Average caching efficiency versus the SINR target threshold of each user.

results of RRH clustering obtained by the ME-BB and separate PDD-based algorithms. Note that in the ME-BB algorithm, we had to limit the maximum time spent by the mixed-integer optimizer MOSEK to 5000 seconds, for otherwise the procedure can become extremely time consuming; the corresponding relative optimality gap (ROG) [50] is on the order of 33% or less.[12] The fronthaul reduction performance of the ME-BB and separete PDD-based algorithms is plotted for different representative cases in Fig. 5. In general, we find that these algorithms are more aggressive, that is, in the final RRH clustering solution, the RRHs tend to serve only one type of users. As noted above, the ME-BB algorithm can only obtain suboptimal solutions in 5000 seconds, and as a result the performance of the separate PDD-based algorithm is better in certain cases with large K. Hence, among the two separate design algorithms, we only provide the results of the separate PDD-based algorithm in the sequel.

In Fig. 6 (a) and (b), we examine the average caching efficiency versus X and Y with fixed $\gamma = 5$dB. For comparison, we also provide the performance of Algorithm 1 when only C_B or C_P is considered and the performance of a heuristic method, where the RRH clustering is simply determined based on the distances between the RRHs and the users. As we can see from Fig. 6 (a), Algorithm 1 and Algorithm 2 exhibit comparable performance, while the performance of the heuristic method and the separate PDD-based algorithm is far worse. This is intuitively plausible since the heuristic method and the separate PDD-based algorithm do not optimize caching placement, RRH clustering and beamforming jointly. Furthermore, only minimizing C_P can still maintain a certain caching efficiency performance when X and Y are relatively small. This is because the fronthaul traffic reduction $C_B(s_{k,n}, c_{f,n})$ changes linearly with $s_{k,n}$ and $c_{f,n}$, respectively. However, in terms of beamforming design, reducing the cooperation among the RRHs (i.e., decreasing X) may lead to significant changes in the total transmission power $C_P(\mathbf{w}_{k,n})$, where the underlying relationship is generally not linear. As a result, when X is large, the performance gain offered by RRH

[12]In the simulations, the final ROG within 5000 seconds is highly dependent to the problem size, i.e., with increasing K, more time is needed to obtain the same optimality gap.

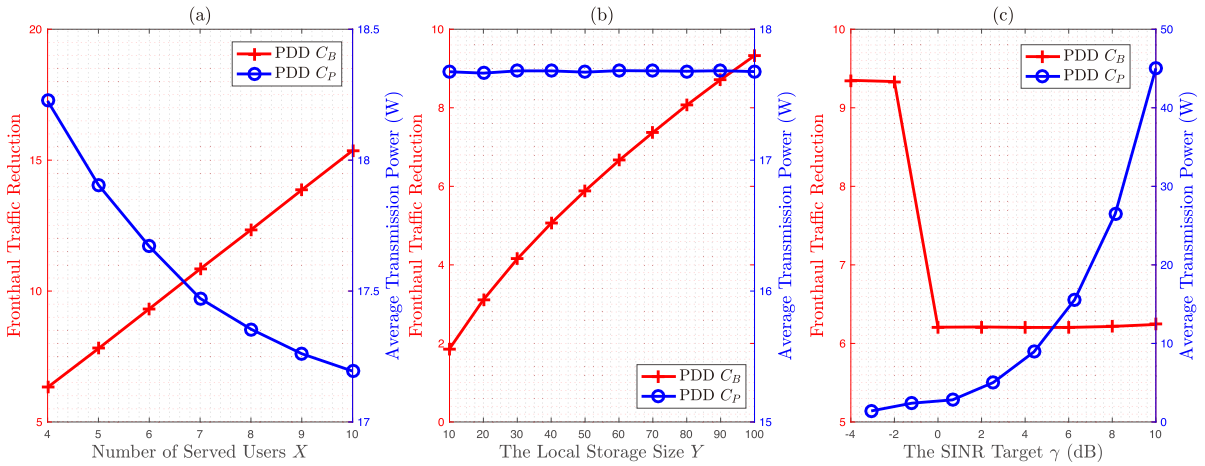


Fig. 8. Fronthaul traffic reduction and total transmission power versus X, Y and γ.

cooperation nearly saturates and thus optimizing content placement is more important. Meanwhile, the performance obtained by only maximizing C_B is poor since the costs of transmission power are ignored. It is also observed that the achieved caching efficiency is in direct proportion to the service capability of the RRHs, as represented by X. Indeed, with increasing X, smarter content placement can be realized, which reduces the fronthaul utilization. Similar trends as those noted with Fig. 6 (a) can be observed in Fig. 6 (b). With the increase of the local storage size in each RRH, more fronthaul reduction can be achieved, which results into higher caching efficiency.

In Fig. 7, the average caching efficiency as a function of the SINR target threshold γ with fixed $Y = 100$ and $X = 6$ is investigated. Similar to the result in Fig. 6, the proposed joint design algorithms exhibit a similar performance, which significantly exceeds that of the heuristic method and the separate PDD-based algorithm. The achieved caching efficiency is in inverse proportion to the SINR target γ of each user. This is because as γ increases, the RRHs need to increase their transmission power to satisfy the more stringent SINR requirements, while the fronthaul reduction barely changes with different γ. Besides, as the SINR target γ decreases, the achieved caching efficiencies of Algorithm 1 (only minimizing C_P), the ME-BB and separate PDD-based algorithms become close to that of the joint design algorithms, since in this case the transmission power is less dominant in achieving high caching efficiency.

Moreover, in Fig. 8, we investigate the achieved C_B and C_P by Algorithm 2 with the same system settings as in Fig. 6 and Fig. 7. Specifically, from Fig. 8 (a), we can see that increasing the number of served users X improves the fronthaul traffic reduction significantly, which is reasonable since a larger X potentially increases the number of 1s in $\{s_{k,n}\}$ and this in turn increases C_B. For the total transmission power, however, the performance gain offered by larger X gradually decreases, because involving some less beneficial RRHs only brings minor performance gain. In Fig. 8 (b), we can observe that the increase of the local storage Y also improves C_B significantly since more files can be cached in each RRH, however C_P merely changes with different values of Y. Finally, in Fig. 8 (c), it can be observed that the required SINR threshold γ has the most profound impact on C_P. In particular, when C_P is larger than some specific value, the proposed PDD-based algorithm will sacrifice C_B for C_P, since this offers better caching efficiency performance.

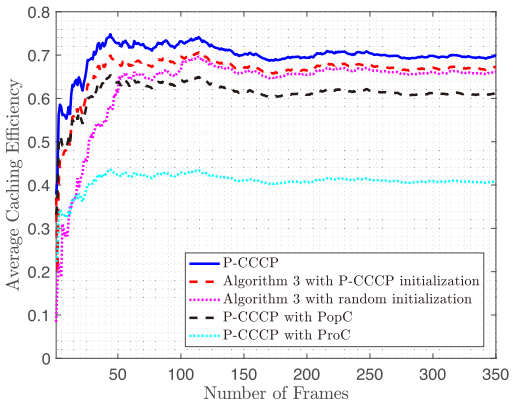


Fig. 9. Average caching efficiency achieved by algorithm 3 versus the number of frames.

Finally, in Fig. 9, we illustrate the performance of the proposed two-timescale joint design algorithm, i.e. Algorithm 3, when $N_s = 50$. For comparison, we also provide the performance obtained by running Algorithm 1 with two fixed caching strategies, i.e., PopC and ProC. It can be seen that the performance of Algorithm 1 can be viewed as an upper bound to that of Algorithm 3. Furthermore, when initialized by Algorithm 1, Algorithm 3 converges much faster than that with random initialization. As Algorithm 1 and Algorithm 2 exhibit similar performance, this also shows the importance and necessity of these single-timescale algorithms. Finally, it can be observed that the proposed Algorithm 3 performs better than Algorithm 1 with fixed caching, which further demonstrates the superiority of the proposed joint design approach. Since the assumed content popularity distribution is non-uniform, therefore, the performance of Algorithm 1 with PopC is better than that with ProC.

VII. Conclusion and Future Work

In this work, we have studied the problem of joint transceiver design for a content-aware C-RAN system.

An optimization framework was presented in which the ratio between fronthaul reduction and transmission power cost, called caching efficiency, is employed as the objective function. By tacking advantage of the Dinkelbach method, two efficient algorithms were proposed to jointly optimize the downlink beamforming vectors, the RRH clustering and the caching placement. To arrive at Algorithm 1 (P-CCCP), we combined the penalty method, the CCCP technique and the BCD method and showed that by introducing auxiliary variables and penalizing certain constraints into the objective function, the original optimization problem could be convexified. In the case of Algorithm 2 (PDD), we utilized the PDD framework and showed that with the introduction of auxiliary variables, the original problem could be addressed by sequentially updating the design variables (in both primal and dual domains) either in closed-form or via the bisection method. Furthermore, a two-timescale joint design algorithm (Algorithm 3) was proposed where the content placement is updated over a larger timescale. We showed that the P-CCCP and PDD-based algorithms can be employed as powerful initialization methods for Algorithm 3. Simulation results were presented, showing that the proposed algorithms exhibit very good performance.

Finally, as motivation for future research, we briefly discuss some important issues/aspects of this work that were not yet addressed or fully unveiled:

1. For simplicity, we only considered the scheduled users, also known as active users, in our study. Generally, the content preferences and RRH clusterings of both active and inactive users will affect the content placement policy. The joint optimization of downlink beamforming, RRH clustering and content placement for both active and inactive users, is more challenging to model and solve. It remains unclear whether the performance gain obtained by considering the impacts of inactive users is sufficiently large to justify the increased complexity, which needs further investigation.

2. In this work, the user preference profile $P_{k,f}$, i.e., the user request probability, was assumed to be known *a prior* by means of proper learning procedures. As a result, in problem (7), the expectation is only taken over the random channel realizations. However, it is more general to assume that the user requests are also random variables. Please see Appendix E for a possible formulation taking this aspect into account.

3. Instead of assuming fixed user preferences for file type, a more general approach would be to consider the scenario where content popularity is dynamically changing. In this case, the historical content requests of the users and user-RRH association information could be exploited to predict the future content popularity for better caching efficiency.

APPENDIX

A. Solution of Problem (27)

The Lagrangian of problem (27) can be formulated as

$$\begin{aligned}\mathcal{L}(\{s_{k,n},\lambda\}) \triangleq & -C_B(s_{k,n},c_{f,n})+\lambda\Big(\sum_{k\in\mathcal{K}} s_{k,n}-X_n\Big)\\ &+\frac{1}{2\rho_m}\sum_{k\in\mathcal{K}}\|(1-s_{k,n})\mathbf{J}_n\mathbf{w}_k+\rho_m\boldsymbol{\xi}_{k,n}^m\|^2\\ &+\frac{1}{2\rho_m}\sum_{k\in\mathcal{K}}\Big((s_{k,n}(1-\hat{s}_{k,n})+\rho_m\lambda_{k,n}^m)^2\\ &+(s_{k,n}-\hat{s}_{k,n}+\rho_m\hat{\lambda}_{k,n}^m)^2\Big). \end{aligned} \tag{37}$$

Taking the derivative with respect to $s_{k,n}$ and equating the result to 0, we obtain

$$s_{k,n}=(a_{k,n}-\lambda\rho_m)/b_{k,n}, \tag{38}$$

where $a_{k,n}=\rho_m\sum_{f\in\mathcal{F}}P_{k,f}c_{f,n}+(\hat{s}_{k,n}-1)\rho_m\lambda_{k,n}^m+(\hat{s}_{k,n}-\rho_m\hat{\lambda}_{k,n}^m)+\frac{\rho_m}{2}\mathbf{w}_k^{kH}\mathbf{J}_n^H\boldsymbol{\xi}_{k,n}^m+\frac{\rho_m}{2}\boldsymbol{\xi}_{k,n}^{mH}\mathbf{J}_n\mathbf{w}_k+\mathbf{w}_k^{kH}\mathbf{J}_n^H\mathbf{J}_n\mathbf{w}_k$ and $b_{k,n}=(1-\hat{s}_{k,n})^2+\mathbf{w}_k^{kH}\mathbf{J}_n^H\mathbf{J}_n\mathbf{w}_k+1$. According to the complementary slackness condition [52] $\lambda\left(\sum_{k\in\mathcal{K}}s_{k,n}-X_n\right)=0$, we consider the following two cases:

- If $\lambda=0$, it follows from (38) that $s_{k,n}=\frac{a_{k,n}}{b_{k,n}}$. Hence, if $\sum_{k\in\mathcal{K}}s_{k,n}\le X_n$ is satisfied, $s_{k,n}=\frac{a_{k,n}}{b_{k,n}}$ is the optimal solution of problem (27).
- If $\lambda>0$, upon substitution of (38) into $\sum_{k\in\mathcal{K}}s_{k,n}=X_n$, we can obtain

$$\lambda=\left(\sum_{k\in\mathcal{K}}\frac{a_{k,n}}{b_{k,n}}-X_n\right)\Big/\sum_{k\in\mathcal{K}}\frac{\rho_m}{b_{k,n}}, \tag{39}$$

  and $s_{k,n}$ can be calculated by substituting (39) into (38), which yields the optimal solution of problem (27).

Therefore, we conclude that the optimal solution of problem (27) can be obtained in closed-form by considering the above two cases.

B. Solution of Problem (28)

We first introduce the following auxiliary variables: $\mathbf{x}_k=\left[(\mathbf{w}_1^k)^H,\cdots,(\mathbf{w}_k^k)^H,\cdots,(\mathbf{w}_K^k)^H\right]^H$, $\mathbf{P}_k=[\mathbf{0}_{NL\times(k-1)NL},\mathbf{I}_{NL\times NL},\mathbf{0}_{NL\times(K-k)NL}]\in\{0,1\}^{NL\times KNL}$, such that $\mathbf{P}_j\mathbf{x}_k=\mathbf{w}_j^k$ holds. We then observe that problem (28) can be equivalently formulated as follows:

$$\begin{aligned}&\min_{\mathbf{x}_k}\ \mathbf{x}_k^H\mathbf{A}_k\mathbf{x}_k+\mathbf{x}_k^H\mathbf{b}_k+\mathbf{c}_k\mathbf{x}_k\\ &\text{s.t. }\mathbf{x}_k^H\mathbf{D}_k\mathbf{x}_k\ge\sigma_k^2,\end{aligned} \tag{40}$$

where

$$\begin{aligned}\mathbf{A}_k\triangleq & \left(\frac{K-1}{2\rho_m}+\eta\right)\mathbf{P}_k^H\mathbf{P}_k+\frac{1}{2\rho_m}\sum_{j\in\mathcal{K}\setminus\{k\}}\mathbf{P}_j^H\mathbf{P}_j\\ &+\frac{1}{2\rho_m}\sum_{n\in\mathcal{N}}(1-s_{k,n})^2\mathbf{P}_k^H\mathbf{J}_n^H\mathbf{J}_n\mathbf{P}_k, \end{aligned} \tag{41}$$

$$\begin{aligned}\mathbf{b}_k\triangleq & \frac{1}{2\rho_m}\sum_{j\in\mathcal{K}\setminus\{k\}}\Big(\mathbf{P}_k^H\left(\rho_m\boldsymbol{\mu}_{j,k}^m-\mathbf{w}_k^j\right)\\ &-\mathbf{P}_j^H\left(\mathbf{w}_j^j+\rho_m\boldsymbol{\mu}_{k,j}^m\right)\Big)\\ &+\frac{1}{2\rho_m}\sum_{n\in\mathcal{N}}(1-s_{k,n})\mathbf{P}_k^H\mathbf{J}_n^H\rho_m\boldsymbol{\xi}_{k,n}^m, \end{aligned} \tag{42}$$

$$\begin{aligned}\mathbf{c}_k\triangleq & \frac{1}{2\rho_m}\sum_{j\in\mathcal{K}\setminus\{k\}}\Big(\left(\rho_m\boldsymbol{\mu}_{j,k}^m-\mathbf{w}_k^j\right)^H\mathbf{P}_k\\ &-\left(\mathbf{w}_j^j+\rho_m\boldsymbol{\mu}_{k,j}^m\right)^H\mathbf{P}_j\Big)\\ &+\frac{1}{2\rho_m}\sum_{n\in\mathcal{N}}\rho_m\boldsymbol{\xi}_{k,n}^m(1-s_{k,n})\mathbf{J}_n\mathbf{P}_k, \end{aligned} \tag{43}$$

$$\mathbf{D}_k \triangleq \frac{1}{\gamma_k}\mathbf{P}_k^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{P}_k - \sum_{j\in\mathcal{K}\setminus\{k\}}\mathbf{P}_j^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{P}_j. \quad (44)$$

Since $\mathbf{A}_k$ is a full-rank matrix, we can decompose it as $\mathbf{A}_k = \mathbf{A}_k^{\frac{1}{2}}\mathbf{A}_k^{\frac{1}{2}}$. Furthermore, by introducing the substitution $\mathbf{y}_k = \mathbf{A}_k^{\frac{1}{2}}\mathbf{x}_k$, problem (40) can be rewritten as

$$\begin{aligned}&\min_{\mathbf{y}_k}\ \mathbf{y}_k^H\mathbf{y}_k + \mathbf{y}_k^H\mathbf{A}_k^{-\frac{1}{2}}\mathbf{b}_k + \mathbf{c}_k\mathbf{A}_k^{-\frac{1}{2}}\mathbf{y}_k\\ &\text{s.t. } \mathbf{y}_k^H\mathbf{A}_k^{-\frac{1}{2}}\mathbf{D}_k\mathbf{A}_k^{-\frac{1}{2}}\mathbf{y}_k \geq \sigma_k^2.\end{aligned} \quad (45)$$

Next, we focus on the optimal solution of problem (45), whose Lagrange function can be expressed as

$$\begin{aligned}\mathcal{L} &= \mathbf{y}_k^H\mathbf{y}_k + \mathbf{y}_k^H\mathbf{A}_k^{-\frac{1}{2}}\mathbf{b}_k + \mathbf{c}_k\mathbf{A}_k^{-\frac{1}{2}}\mathbf{y}_k\\ &+\lambda_k\big(\sigma_k^2 - \mathbf{y}_k^H\mathbf{A}_k^{-\frac{1}{2}}\mathbf{D}_k\mathbf{A}_k^{-\frac{1}{2}}\mathbf{y}_k\big),\end{aligned} \quad (46)$$

where λ_k denotes the dual variable. Employing the eigenvalue decomposition, we can write $\mathbf{A}_k^{-\frac{1}{2}}\mathbf{D}_k\mathbf{A}_k^{-\frac{1}{2}} = \mathbf{V}\mathbf{S}\mathbf{V}^{-1}$, where $\mathbf{V}$ is unitary and $\mathbf{S} = \text{diag}(s_1, \cdots, s_{KNL})$ is diagonal. Note that in order for the problem to be feasible, the dual variable should satisfy $\mathbf{I} - \lambda_k\mathbf{V}\mathbf{S}\mathbf{V}^{-1} \succeq \mathbf{0}$, which is equivalent to $\mathbf{I} - \lambda_k\mathbf{S} \succeq \mathbf{0}$. Taking the derivative of $\mathcal{L}$ with respect to $\mathbf{y}_k^*$, we obtain

$$\mathbf{y}_k + \mathbf{A}_k^{-\frac{1}{2}}\mathbf{b}_k - \lambda_k\mathbf{V}\mathbf{S}\mathbf{V}^{-1}\mathbf{y}_k = \mathbf{0}, \quad (47)$$

which is equivalent to

$$\mathbf{y}_k = \mathbf{V}(\mathbf{I} - \lambda_k\mathbf{S})^{-1}\mathbf{V}^{-1}(-\mathbf{A}_k^{-\frac{1}{2}}\mathbf{b}_k). \quad (48)$$

The Lagrange dual variable λ_k can be obtained by the bisection method, and the corresponding upper bound $\overline{\lambda}_k$ and lower bound $\underline{\lambda}_k$ can be found by resorting to $\mathbf{I} - \lambda_k\mathbf{S} \succeq 0$, which results into $\overline{\lambda}_k = \frac{1}{\max(0,\, s_1,\, \cdots,\, s_{KNL})}$ and $\underline{\lambda}_k = 0$.

C. Solution of Problem (29)

The Lagrange function of problem (29) can be expressed as

$$\begin{aligned}\mathcal{L} &= \tfrac{1}{2\rho_m}(s_{k,n}(1-\hat{s}_{k,n}) + \rho_m\lambda_{k,n}^m)^2\\ &+\tfrac{1}{2\rho_m}(s_{k,n} - \hat{s}_{k,n} + \rho_m\hat{\lambda}_{k,n}^m)^2 + \lambda(\hat{s}_{k,n}-1) - \mu\hat{s}_{k,n},\end{aligned} \quad (49)$$

where λ and μ denote the dual variables corresponding to the constraints $\hat{s}_{k,n} \leq 1$ and $0 \leq \hat{s}_{k,n}$, respectively. Taking the derivative with respect to $\hat{s}_{k,n}$, we obtain

$$\hat{s}_{k,n} = a_{k,n}/b_{k,n}, \quad (50)$$

where $a_{k,n} = s_{k,n}(s_{k,n} + \rho_m\lambda_{k,n}^m) + (s_{k,n} + \rho_m\hat{\lambda}_{k,n}^m) - \lambda\rho_m + \mu\rho_m$ and $b_{k,n} = s_{k,n}^2 + 1$. Next, we consider the following three cases:

- $\lambda = 0$ and $\mu = 0$: if (50) satisfies $0 \leq \hat{s}_{k,n} \leq 1$, then this is the optimal solution;
- $\lambda = 0$ and $\mu > 0$: we have $\hat{s}_{k,n} = 0$, then if $\mu = -\frac{s_{k,n}^2 + s_{k,n}}{\rho_m} - s_{k,n}\lambda_{k,n}^m - \hat{\lambda}_{k,n}^m > 0$ holds, $\hat{s}_{k,n} = 0$ is the optimal solution;
- $\lambda > 0$ and $\mu = 0$: we have $\hat{s}_{k,n} = 1$, then if $\lambda = \frac{s_{k,n}-1}{\rho_m} + s_{k,n}\lambda_{k,n}^m + \hat{\lambda}_{k,n}^m > 0$ holds, $\hat{s}_{k,n} = 1$ is the optimal solution.

Therefore, the optimal solution of problem (29) can be obtained in closed-form.

D. A Simple Initialization Method

In this appendix, we propose a simple initialization method based on ZF beamforming. We first address the initialization of the beamforming vectors $\{\mathbf{w}_k\}$ by considering the following ZF problem:

$$\begin{aligned}&\min_{\mathbf{w}_k}\ \|\mathbf{w}_k\|^2\\ &\text{s.t. } |\mathbf{h}_k^H\mathbf{w}_k|^2 \geq \gamma_k\sigma_k^2,\ \mathbf{H}_k^H\mathbf{w}_k = \mathbf{0},\end{aligned} \quad (51)$$

where it can be easily seen that the first constraint must be satisfied with equality at optimality. Therefore, we have the following equivalent problem:

$$\begin{aligned}&\min_{\mathbf{w}_k}\ \|\mathbf{w}_k\|^2\\ &\text{s.t. } |\mathbf{h}_k^H\mathbf{w}_k|^2 = \gamma_k\sigma_k^2,\ \mathbf{H}_k^H\mathbf{w}_k = \mathbf{0}.\end{aligned} \quad (52)$$

Defining $\mathbf{w}_k = \sqrt{p_k}\bar{\mathbf{w}}_k$ with $\|\bar{\mathbf{w}}_k\| = 1$, it can be observed that problem (52) is equivalent to

$$\begin{aligned}&\min_{p_k,\,\bar{\mathbf{w}}_k}\ p_k\\ &\text{s.t. } p_k|\mathbf{h}_k^H\bar{\mathbf{w}}_k|^2 = \sigma_k^2\gamma_k,\ \mathbf{H}_k^H\bar{\mathbf{w}}_k = \mathbf{0}.\end{aligned} \quad (53)$$

To achieve the minimum p_k, the optimal $\bar{\mathbf{w}}_k$ should be the optimal solution to the following problem:

$$\begin{aligned}&\max_{\bar{\mathbf{w}}_k}\ |\mathbf{h}_k^H\bar{\mathbf{w}}_k|^2\\ &\text{s.t. } \mathbf{H}_k^H\bar{\mathbf{w}}_k = \mathbf{0},\ \|\bar{\mathbf{w}}_k\| = 1.\end{aligned} \quad (54)$$

Then, it can be easily seen that the optimal solution of problem (54) is given by

$$\bar{\mathbf{w}}_k = \mathbf{U}_k\mathbf{U}_k^H\mathbf{h}_k/\|\mathbf{U}_k\mathbf{U}_k^H\mathbf{h}_k\|, \quad (55)$$

where $\mathbf{U}_k$ denotes the orthogonal basis for the null space of $\mathbf{H}_k^H$. Hence, the optimal p_k is given by $\sigma_k^2\gamma_k/|\mathbf{h}_k^H\bar{\mathbf{w}}_k|^2$.

E. A Possible Formulation With Random Content Request

Let Z_k denote the random variable which represents the content request of user k, whose sample space is $\mathcal{F}$, i.e., $P_{k,f} = P_r(Z_k = f)$ denotes the probability that user k requests content file f. Then, when the user requests are considered as random variables, the total fronthaul traffic reduction of the considered cache-enabled C-RAN would become

$$C_B(s_{k,n}, c_{f,n}) = \sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}} s_{k,n}\sum_{f\in\mathcal{F}}\mathbf{1}_f(Z_k)c_{f,n}, \quad (56)$$

where $\mathbf{1}_f(\cdot)$ is an indicator function defined as

$$\mathbf{1}_f(Z_k) = \begin{cases}1, \text{ if } Z_k = f,\\ 0, \text{ if otherwise.}\end{cases} \quad (57)$$

As a result, we can consider the following two-timescale stochastic optimization problem:

$$\begin{aligned}\min_{\{\mathbf{w}_{k,n}s_{k,n}c_{f,n}\}}\ & f(c_{f,n},\boldsymbol{\Theta}) \triangleq \mathbb{E}_{\{\mathbf{h}_k,Z_k\}}(-C(s_{k,n}, c_{f,n}, \mathbf{w}_{k,n}))\\ \text{s.t. } & \text{(6b)} - \text{(6f)},\end{aligned} \quad (58)$$

which is worthy of further investigation.

REFERENCES

[1] M.-M. Zhao, Y. Cai, M.-J. Zhao, and B. Champagne, "Joint content placement, RRH clustering and beamforming for cache-enabled cloud-RAN," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[2] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[3] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[4] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[5] H. Liu, Z. Chen, X. Tian, X. Wang, and M. Tao, "On content-centric wireless delivery networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 118–125, Dec. 2014.

[6] A. Liu and V. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.

[7] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 41, Feb. 2015.

[8] M. Ji, G. Caire, and A. F. Molisch, "Wireless Device-to-Device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[9] "C-RAN: the road towards green RAN," China Mobile, Hong Kong, White Paper, Version 3.0, Dec. 2013.

[10] M. Hong, R.-Y. Sun, H. Baligh, and Z.-Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogenous networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 226–240, Feb. 2013.

[11] 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.

[12] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.

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

[14] B. Hu, C. Hua, J. Zhang, C. Chen, and X. Guan, "Joint fronthaul multicast beamforming and user-centric clustering in downlink C-RANs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5395–5409, Aug. 2017.

[15] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.

[16] A. Liu and V. K. N. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.

[17] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2014, pp. 1370–1374.

[18] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Jun. 2015, pp. 809–813.

[19] M. Ali Maddah-Ali and U. Niesen, "Cache-aided interference channels," 2015, *arXiv:1510.06121*. [Online]. Available: http://arxiv.org/abs/1510.06121

[20] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[21] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Content caching at the wireless network edge: A distributed algorithm via belief propagation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[22] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[23] B. Dai and W. Yu, "Joint user association and content placement for cache-enabled wireless access networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 3521–3525.

[24] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct. 2017.

[25] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.

[26] W. Han, A. Liu, and V. K. N. Lau, "PHY-caching in 5G wireless networks: Design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.

[27] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[28] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, May 2015.

[29] G. R. Lanckriet and B. K. Sriperumbudur, "On the convergence of the concave-convex procedure," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1759–1767.

[30] Y. Cai, Q. Shi, B. Champagne, and G. Y. Li, "Joint transceiver design for secure downlink communications over an Amplify-and-Forward MIMO relay," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3691–3704, Sep. 2017.

[31] Q. Shi, M. Hong, X. Fu, and T.-H. Chang, "Penalty dual decomposition method for nonsmooth nonconvex optimization," 2017, *arXiv:1712.04767*. [Online]. Available: http://arxiv.org/abs/1712.04767

[32] A. Liu, V. K. N. Lau, and M.-J. Zhao, "Online successive convex approximation for two-stage stochastic nonconvex optimization," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5941–5955, Nov. 2018.

[33] C. Yang, Z. Chen, B. Xia, and J. Wang, "When ICN meets C-RAN for HetNets: An SDN approach," *IEEE Commun. Mag.*, vol. 53, no. 11, pp. 118–125, Nov. 2015.

[34] P. Blasco and D. Gunduz, "Learning-based optimization of cache content in a small cell base station," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 1897–1903.

[35] E. Bastug, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," in *Proc. 13th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2015, pp. 161–166.

[36] B. N. Bharath, K. G. Nagananda, and H. V. Poor, "A learning-based approach to caching in heterogenous small cell networks," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1674–1686, Apr. 2016.

[37] W.-C. Liao, M. Hong, Y.-F. Liu, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3939–3952, Aug. 2014.

[38] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx

[39] K.-Y. Wang, A. M.-C. So, T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: Tractable approximations by conic optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5690–5705, Nov. 2014.

[40] M. R. Hestenes, "Multiplier and gradient methods," *J. Optim. Theory Appl.*, vol. 4, no. 5, pp. 303–320, Nov. 1969.

[41] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," in *Optimization*, R. Fletcher, Ed. New York, NY, USA: Academic, 1972, pp. 283–298.

[42] M.-M. Zhao, Q. Shi, Y. Cai, M.-J. Zhao, and Q. Yu, "Decoding binary linear codes using penalty dual decomposition method," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 958–962, Jun. 2019.

[43] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

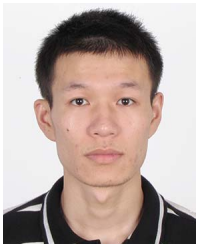
[44] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.

[45] M.-M. Zhao, Q. Shi, Y. Cai, and M.-J. Zhao, "Joint transceiver design for full-duplex cloud radio access networks with SWIPT," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5644–5658, Sep. 2017.

[46] F. Rinaldi, F. Schoen, and M. Sciandrone, "Concave programming for minimizing the zero-norm over polyhedral sets," *Comput. Optim. Appl.*, vol. 46, no. 3, pp. 467–486, Jul. 2010.

[47] R. Sun, Y. Wang, N. Cheng, H. Zhou, and X. Shen, "QoE driven BS clustering and multicast beamforming in cache-enabled C-RANs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[48] Y. Li, M. Xia, and Y.-C. Wu, "First-order algorithm for content-centric sparse multicast beamforming in large-scale C-RAN," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5959–5974, Sep. 2018.
[49] A. H. Land and A. G. Doig, "An automatic method of solving discrete programming problems," *Econometrica*, vol. 28, no. 3, p. 497, Jul. 1960.
[50] *The MOSEK Modeling Cookbook*, MOSEK ApS, Copenhagen, Denmark, 2012.
[51] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE Conf. Comput. Commun. 18th Annu. Joint Conf. IEEE Comput. Commun. Societies. Future (INFOCOM)*, vol. 1, Mar. 1999, pp. 126–134.
[52] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.



Ming-Min Zhao received the B.Eng. and Ph.D. degrees in information and communication engineering from Zhejiang University in 2012 and 2017, respectively. From December 2015 to August 2016, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. From July 2017 to July 2018, he worked as a Research Engineer with Huawei Technologies Company, Ltd. Since May 2019, he has been a Visiting Scholar with the Department of Electrical and Computer Engineering, National University of Singapore. He is currently a Lecturer with the College of Information Science and Electronic Engineering, Zhejiang University. His research interests include channel coding, algorithm design and analysis for advanced MIMO, cooperative communication, and machine learning for wireless communications.

Yunlong Cai (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from the University of York, York, U.K., in 2010. Since February 2011, he has been with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, where he is currently a Full Professor. From August 2016 to January 2017, he was a Visiting Scholar with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. His research interests include transceiver design for multiple-antenna systems, mmWave communications, full-duplex communications, UAV communications, cooperative communications, and machine learning for communications. He is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.

Min-Jian Zhao (Member, IEEE) received the M.Sc. and Ph.D. degrees in communication and information systems from Zhejiang University, Hangzhou, China, in 2000 and 2003, respectively.

He is currently a Professor with the Department of Information Science and Electronic Engineering, Zhejiang University. His research interests include modulation theory, channel estimation and equalization, and signal processing for wireless communications.

Benoit Champagne (Senior Member, IEEE) received the B.Ing. degree in engineering physics from the École Polytechnique de Montréal in 1983, the M.Sc. degree in physics from the Université de Montréal in 1985, and the Ph.D. degree in electrical engineering from the University of Toronto in 1990.

From 1990 to 1999, he was an Assistant and then an Associate Professor with INRS-Telecommunications, Université du Quebec, Montréal. In1999, he joined McGill University, Montreal, where he is currently a Full Professor with the Department of Electrical and Computer Engineering. He has served as the Associate Chairman of graduate studies with the Department from 2004 to 2007. His research focuses on the study of advanced algorithms for the processing of communication signals by digital means. His interests span many areas of statistical signal processing, including detection and estimation, sensor array processing, adaptive filtering, and applications thereof to broadband communications and audio processing, where he has coauthored more than 250 referred publications. His research has been funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Fonds de Recherche sur la Nature et les Technologies from the Government of Quebec, as well as some major industrial sponsors, including Nortel Networks, Bell Canada, InterDigital, and Microsemi. He has been an Associate Editor for the *EURASIP Journal on Applied Signal Processing* from 2005 to 2007, the IEEE SIGNAL PROCESSING LETTERS from 2006 to 2008, and the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2010 to 2012, as well as a Guest Editor for two special issues of the *EURASIP Journal on Applied Signal Processing* published in 2007 and 2014, respectively. He has also served on the Technical Committees of several international conferences in the fields of communications and signal processing. In particular, he was the Registration Chair of the IEEE ICASSP 2004, the Co-Chair, Antenna and Propagation Track, of IEEE VTC–Fall 2004, the Co-Chair, Wide Area Cellular Communications Track, of the IEEE PIMRC 2011, the Co-Chair, Workshop on D2D Communications, of the IEEE ICC 2015, and the Publicity Chair of the IEEE VTC–Fall 2016.

Theodoros A. Tsiftsis (Senior Member, IEEE) was born in Lamia, Greece, in 1970. He received the B.Sc. degree in physics from the Aristotle University of Thessaloniki, Greece, in 1993, the M.Sc. degree in digital systems engineering from Heriot-Watt University, Edinburgh, U.K., in 1995, the M.Sc. degree in decision sciences from the Athens University of Economics and Business, in 2000, and the Ph.D. degree in electrical engineering from the University of Patras, Greece, in 2006.

He is currently a Professor with the School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai, China, and also an Honorary Professor with Shandong Jiaotong University, Jinan, China. His research interests include the broad areas of cognitive radio, communication theory, wireless powered communication systems, optical wireless communication, and ultrareliable low-latency communication. He has served as a Senior or Associate Editor in the editorial boards of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE COMMUNICATIONS LETTERS, *IET Communications*, and the *IEICE Transactions on Communications*. He is also an Area Editor of Wireless Communications II of the IEEE TRANSACTIONS ON COMMUNICATIONS and an Associate Editor of the IEEE TRANSACTIONS ON MOBILE COMPUTING. He has been appointed to a two-year term as an IEEE Vehicular Technology Society Distinguished Lecturer (IEEE VTS DL), in 2018.