Mobile Edge Computing Meets mmWave Communications: Joint Beamforming and Resource Allocation for System Delay Minimization

Cunzhuo Zhao, Yunlong Cai, An Liu, Minjian Zhao, and Lajos Hanzo

***Abstract*—Mobile edge computing (MEC) has been identified as a key technique of next-generation wireless networks, which supports cloud computing along with other compelling service capabilities at the network's edge with the objective of reducing the system delay. As one of the prospective candidates for new spectrum in next-generation networks, millimeter wave (mmWave) communications has been gaining significant attention as a benefit of its high rate. Hence we conceive a joint hybrid beamforming and resource allocation algorithm for mmWave MEC. Explicitly, we jointly optimize the analog beamforming vectors at the users, the analog and digital beamforming matrices at the base station (BS), the computation task offloading ratios and resource allocation at the MEC server for minimizing the maximum system delay subject to the affordable communication and computing budget. We conceive a powerful algorithm for solving this challenging nonconvex optimization problem with coupled constraints based on the penalty dual decomposition (PDD) technique. The proposed algorithm can be implemented in a parallel and distributed fashion. Our numerical results demonstrate the superiority of the proposed algorithm by quantifying the benefits of intrinsically amalgamating MEC with mmWave communications.**

***Index Terms*—Millimeter wave, hybrid beamforming, mobile edge computing, resource allocation, distributed implementation**

I. Introduction

Mobile edge computing (MEC) has been identified as a key technique of supporting cloud computing and other compelling new services at the network edge [1]. More particularly, the MEC servers are connected to the base stations (BSs) through a backhaul link or are directly installed at the BSs using generic computing platforms for providing cloud-computing services in close proximity of mobile users [2]. The MEC systems have the advantage of significantly reducing the system delay, whilst avoiding tele-traffic congestion. A number of flawless multimedia services, such as augmented reality, caching, surveillance and security have exploited MEC by assigning their latency-sensitive applications to the MEC servers for cloud-based execution [3].

When aiming for low system latency and energy-efficient MEC, jointly designing radio- and computational resource management is promising. In [4]–[7], the authors studied the joint optimization of the transmission power, CPU frequency and task offloading ratio in a single-user MEC system in order to minimize the user's time delay subject to the energy budget. Furthermore, the corresponding multi-user scenarios have been considered in [8]–[12]. Specifically, in [8], the optimal resource allocation and offloading strategy are developed for minimizing the total energy consumption under a computational delay constraint. In [9], the joint allocation of a user's transmit power and computational resources are investigated, where the authors have adopted a decomposition technique for optimizing the offloading decisions and resource allocation sequentially. The authors of [10] considered a multi-user time division multiple access (TDMA) video compression offloading scheme relying both on cooperative communication and on computational resource allocation, which minimizes the execution delay of both local compression, of edge compression and partial compression offloading schemes. The study of [11] employed the classic Lyapunov optimization based stochastic task arrival model for striking an energy and system delay tradeoff. A decentralized game-theoretic computation offloading algorithm is proposed for multi-channel environments in [12].

As one of the prospective candidates for next-generation wireless communication systems, millimeter wave (mmWave) techniques have been proposed, which have a potential bandwidth of upto 10 GHz and a transmission rate of upto 20 Gbits/s [13]–[15]. As a benefit of its substantial bandwidth and high transmission rate, mmWave techniques are eminently suitable for MEC systems. This idea also meets the vision of the 5G-MiEdge initiative (Millimeter-wave Edge Cloud as an Enabler for 5G Ecosystem) for the forthcoming Tokyo 2020 Olympics [16]. However, compared to the regular sub-5 GHz frequency band, mmWave channels impose a high path-loss, high penetration loss, rain-attenuation, etc [17]. Fortunately, given their mm-scale wave-length, large antenna arrays can be accommodated in a compact space for high-gain directional beamforming. However, in contrast to traditional transceivers, the high cost and high energy consumption of the analog-to-digital converters as well as of the radio frequency (RF) chains render fully digital processing in mmWave systems unfeasible. Although fully analog beamforming is cost-effective, since it only relies on analog phase shifters, the advantage of analog beamforming comes at the cost of dealing with a

The work of Y. Cai was supported in part by the National Natural Science Foundation of China under Grants 61831004 and 61971376, and in part by the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under Grant LR19F010002. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/Noo4558/1, EP/PO34284/1, COALESCE, of the Royal Society's Global Challenges Research Fund Grant as well as of the European Research Council's Advanced Fellow Grant QuantCom. (Correspondence authors: Yunlong Cai; Minjian Zhao.)

C. Zhao, Y. Cai, A. Liu and M. Zhao are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (Email: zhaocz@zju.edu.cn; ylcai@zju.edu.cn; anliu@zju.edu.cn; mjzhao@zju.edu.cn).

L. Hanzo is with the Department of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K. (Email: lh@ecs.soton.ac.uk).

single data stream, which limits the signal processing and multiplexing capability of the system [18]. To circumvent this limitation, a hybrid beamforming architecture, consisting of baseband digital and RF analog beamformers, has been widely adopted by mmWave systems [19], [20]. Some representative algorithms have been proposed in [21]–[24] for the design of hybrid beamforming. In [21], the authors have exploited the sparse structure of the mmWave channel impulse response (CIR) to provide an algorithmic precoding solution based on the concept of orthogonal matching pursuit (OMP). As an extension of the OMP, an analog beamforming algorithm based on channel matching (CM) has been proposed in [22]. The authors of [23] developed a hybrid transceiver design based on manifold optimization combined with antenna selection for massive MIMO mmWave systems. A heuristic iterative algorithm is proposed in [24] for designing the hybrid beamformer relying on quantized phase shifters.

Let us now briefly consider the integration of MEC and mmWave techniques, which has substantial promise. A beneficial application scenario corresponds to public surveillance and high definition (HD) video broadcast services, where large amounts of online monitoring data gleaned from access points can be promptly analyzed by a MEC server [25]. In automated driving, the integration of MEC and mmWave techniques can also be beneficially exploited, where the information generated through sensors is analyzed in support of safe maneuvering [25]. In the existing literatures, although the authors of [26] combined MEC with mmWave communications in a single-user multi-link scenario, to the best of our knowledge, the systematic design of multi-user mmWave MEC systems has not been investigated in the open literature. To this end, we are motivated to investigate the delay minimization problem in this system. We propose a joint hybrid beamforming and resource allocation algorithm for a multi-user mmWave MEC system, where the users are capable of offloading their latency-sensitive tasks to the BS equipped with the MEC server. We seek to jointly optimize the analog beamforming vectors at the users, the analog and digital beamforming matrices at the BS, the task offloading ratios, and the resource allocation at the MEC server in order to minimize the maximum system delay subject to the communication and computing budget. It is quite a challenge to globally solve the resultant optimization problem, due to the highly nonlinear and nonconvex nature of objective function, as well as owing to the highly coupled constraints.

The main contributions of this treatise are summarized as follows.

1) We first introduce our multi-user mmWave MEC system model, followed by formulating our joint beamforming and resource allocation problem for system delay minimization under specific communication and computing constraints.
2) We then recast this nonlinear and nonconvex optimization problem into an equivalent but more tractable form. We continue by proposing a joint optimization algorithm for the resultant problem based on the penalty dual decomposition (PDD) framework of [27]–[30]. The proposed algorithm ensures convergence to the set of stationary solutions of the original optimization problem. We also show that the proposed algorithm can be implemented in a parallel and distributed fashion. Furthermore, its computational complexity is analyzed.
3) To characterize the benefits of this system, we provide exhaustive simulation results for a range of pertinent system settings. The results clearly demonstrate the superiority of the proposed algorithm and the quantitative benefits of combining MEC with mmWave communications.

The rest of this paper is organized as follows. In Section II, we present our system model of a multi-user mmWave MEC system and formulate the problem considered. We develop our PDD-based joint beamforming design and resource allocation algorithm to solve the resultant problem in Section III. In Section IV, the practical implementation of the proposed algorithm and its computational complexity are discussed. In Section V, simulations are provided and the benefits of the proposed algorithm in the new system are demonstrated. Finally, we offer our conclusions in Section VI.

II. System model and problem formulation

We consider a narrowband mmWave based MEC uplink multi-user multi-input multi-output (MIMO) system. The system contains K users, each of which is equipped with $N_{\rm user} > 1$ antennas and a single RF chain. All the users communicate with a base station (BS) equipped with $N_{\rm BS}$ antennas and $N_{\rm s}$ RF chains ($N_{\rm BS} \geq N_{\rm s}$), connecting to the MEC server through a high speed backhaul link. Each user sends a single data stream modulated by an analog beamforming vector to the BS, and the BS decodes multiple data streams for those K users from the received data vector. To insure spatial multiplexing gain for the K users, we set $K \leq N_{\rm s}$. All the channels between the BS and the users are flat fading.

A. Communication Model

As illustrated in Fig. 1, the transmitted signal from user $k \in \mathcal{K} \triangleq \{1, \cdots, K\}$ can be expressed as $\mathbf{x}_k = \mathbf{v}_{RF,k} s_k$, where $\mathbf{v}_{RF,k} \in \mathbb{C}^{N_{\rm user} \times 1}$ and $s_k \sim \mathcal{CN}(0, P)$ are the analog beamforming vector and data symbol, respectively.

Through a narrowband blocking-fading channel, the received signal of user k after the analog and baseband processing at the BS can be expressed as

$$\hat{s}_k = \mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_k \mathbf{x}_k + \mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \sum_{k' \neq k} \mathbf{H}_{k'} \mathbf{x}_{k'} + \mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{n}_k, \tag{1}$$

where $\mathbf{U}_{RF} \in \mathbb{C}^{N_{\rm BS} \times N_{\rm s}}$ denotes the BS analog beamforming matrix, $\mathbf{U}_{BB} = [\mathbf{u}_{BB,1}, ..., \mathbf{u}_{BB,K}] \in \mathbb{C}^{N_{\rm s} \times K}$ denotes the BS digital beamforming matrix, $\mathbf{H}_k \in \mathbb{C}^{N_{\rm BS} \times N_{\rm user}}$ represents the mmWave channel matrix from user k to the BS, and $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_{\rm BS}})$ is the additive white Gaussian noise at the BS.

For such a system with hybrid beamforming architecture, the throughput of user k (in bps/Hz) can be written as

$$\log_2 \left(1 + \frac{P |\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_k \mathbf{v}_{RF,k}|^2}{\sum_{k' \neq k} P |\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_{k'} \mathbf{v}_{RF,k'}|^2 + \|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H\|^2 \sigma^2} \right). \tag{2}$$

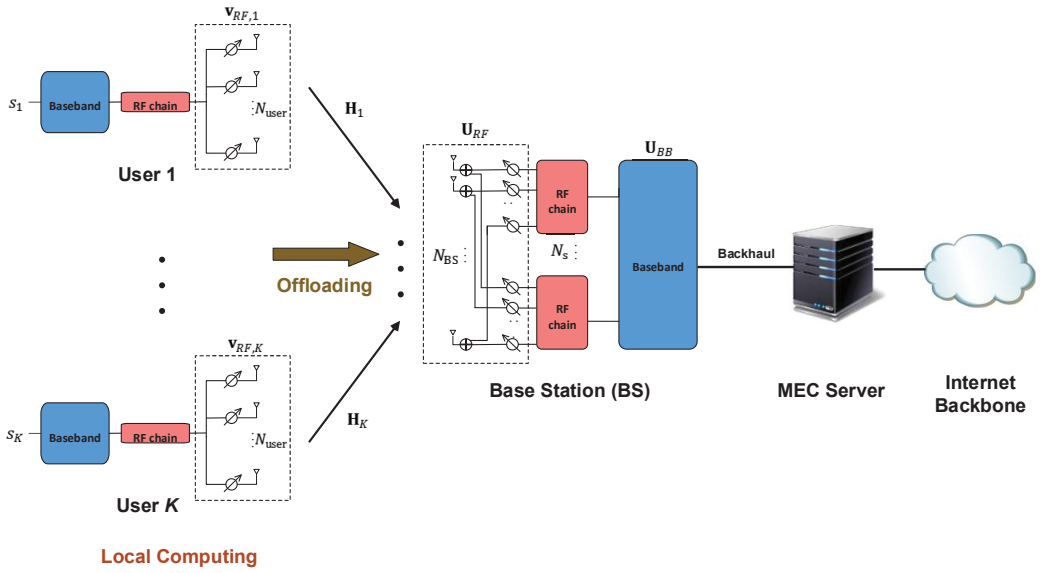

Fig. 1: Illustration of mmWave based MEC network.

B. Computation Model

We assume that there are L_k bits task to be processed for user k, and the tasks can be partially offloaded. The users and the BS are both equipped with computational resources. To further reduce the system delay, the computational resources at both users and the BS can be utilized. Therefore, the tasks can be divided into two parts: one portion is processed locally while the rest is offloaded to the BS and computed at the MEC server. In this work, the size of the computation result is assumed to be quite small, thus we can neglect the feedback delay [6], [7]. Let $\tau_k \in [0,1]$ denote the task offloading ratio for user k, i.e., $\tau_k L_k$ bits are offloaded to the BS and processed by the MEC server, while $(1-\tau_k)L_k$ bits are processed at user k locally. Then the total partial computation offloading time at the edge contains two parts: the raw data transmission time from the user to the BS and the data processing time at the MEC server. Let $C_{1,k}$ represent the computational resources allocated to user k at the MEC server, which satisfies $\sum_{k=1}^{K} C_{1,k} \leq C_{\max}$, where $C_{\max}$ denotes the computational budget at the MEC server, and let $C_{2,k}$ be the local computational capability at the user's end. Naturally, the computational capability depends both on the CPU's clock frequency and on the nature of the tasks, which is best characterized in a normalized form, i.e. by the number of CPU cycles per data bit. Typically we have $C_{2,k} \ll C_{\max}$.

For user k, defining $D_{1,k}$ as the total time for partial offloading and $D_{2,k}$ as the time required for local computation, respectively. Then we have (3) and (4),

$$D_{2,k} \triangleq \frac{(1-\tau_k)L_k}{C_{2,k}}, \tag{4}$$

where W is the bandwidth of the mmWave channel.

For each user, the process of partial offloading and local computation take place simultaneously, so the time required by each user is strongly dominated by the longer delays. We address the question of how to allocate the computational resources to each user and optimize the transmission/reception beamformers at the users/BS so that the tasks of all the users can be processed in the most efficient way.

C. Problem Formulation

In the proposed mmWave MEC system, we concentrate our attention on the joint design of the analog beamforming vectors $\mathbf{v}_{RF,k}$ for user k, of the analog beamforming matrix $\mathbf{U}_{RF}$ and of the digital beamforming matrix $\mathbf{U}_{BB}$ at the BS, as well as the offloading ratio τ_k and the computational resources $C_{1,k}$ allocated to user k in order to minimize the maximum delay among all the users. The system delay minimization problem can be mathematically formulated as

$$\min_{\{\mathbf{v}_{RF,k},\mathbf{U}_{BB},\mathbf{U}_{RF},\tau_k,C_{1,k}\}} \max_k \ \max\left\{D_{1,k}, D_{2,k}\right\} \tag{5a}$$

$$\text{s.t.} \quad |\mathbf{v}_{RF,k}(i)| = 1, \ \forall k \in \mathcal{K}; \ i = 1, \cdots, N_{\text{user}} \tag{5b}$$

$$|\mathbf{U}_{RF}(i,j)| = 1, \ \forall i = 1, \cdots, N_{\text{BS}}; j = 1, \cdots, N_{\text{s}} \tag{5c}$$

$$0 \leq \tau_k \leq 1, \ \forall k \in \mathcal{K} \tag{5d}$$

$$\sum_{k=1}^{K} C_{1,k} \leq C_{\max}. \tag{5e}$$

where constraints (5b) and (5c) impose the unit norm constraints on analog transmit and receive beamforming matrices. Constraint (5e) reflects the computational resource budget at the MEC server.

Note that problem (5) is highly nonconvex and challenging to handle, mainly due to the coupling variables in the objective function and the unit-norm constraints. In the next section, we first conduct a series of transformations for (5) and then propose an efficient joint hybrid beamforming and resource allocation algorithm which ensures convergence to a local stationary solution of the problem.

III. Proposed Joint Beamforming and Resource Allocation Algorithm

In this section, we first convert problem (5) into an equivalent yet more tractable form. By applying the PDD

$$D_{1,k} \triangleq \frac{\tau_k L_k}{C_{1,k}} + \frac{\tau_k L_k}{W \log_2\left(1+\frac{P|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_k \mathbf{v}_{RF,k}|^2}{\sum_{k'\neq k} P|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_{k'} \mathbf{v}_{RF,k'}|^2 + \|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H\|^2 \sigma^2}\right)}, \tag{3}$$

method a set of auxiliary variables and equality constraints are introduced to handle the highly coupled terms, then the problem results in a simpler problem with a number of separable equality constraints. The introduced equality constraints are penalized and dualized into the objective function to formulate the augmented Lagrangian (AL) problem [31], [32]. The proposed PDD-based algorithm has two iteration loops. In the inner loop, we develop an efficient concave-convex procedure (CCCP)-based algorithm [33], [34] to solve the AL problem in a BCD fashion, and in the outer loop we adjust the dual variables or penalty parameter in terms of the constraint violation. Finally we summarize the proposed algorithm and evaluate its computational complexity.

A. Reformulation of Problem (5)

To deal with the nonconvexity of the objective function, we first equivalently convert the optimization problem (5) into a more tractable form. The difficulties lies in the coupling terms in the objective function, and this can be dealt with by introducing auxiliary variables and additional equality constraints. Here based on our experience, we list the guidelines on how to introduce auxiliary variables and equality constraints.

1) Coupling terms should not be contained in the new constraints ;
2) The variables in the same constraint should be jointly optimized;
3) Each variable cannot appear in more than one constraint.

To be specific, guideline 1) ensures that the resulting algorithm converges to a stationary solution of the original problem, while the conventional alternating optimization methods with coupling constraints cannot meet this condition [35], guideline 2) prevents that the updating algorithm gets trapped in a deadlock while updating variables, and guideline 3) helps us to decompose the original problem into a set of subproblems, which could be easily dealt with. Then we first introduce the following auxiliary variables t and $\{t_k\}$, $\{z_{1,k}\}$, $\{z_{2,k}\}, \forall k \in \mathcal{K}$, with the equality constraints $t = t_1 = ... = t_K$. Therefore problem (5) can be rewritten as

$$\min_{\mathcal{Z}} \quad t \tag{6a}$$

$$\text{s.t.} \quad \frac{\tau_k L_k}{C_{1,k}} \leq z_{1,k}, \quad \forall k \tag{6b}$$

$$\frac{\tau_k L_k}{W \log_2\left(1+\frac{P|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_k \mathbf{v}_{RF,k}|^2}{\sum_{k'\neq k} P|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_{k'} \mathbf{v}_{RF,k'}|^2 + \|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H\|^2 \sigma^2}\right)} \leq z_{2,k}, \quad \forall k \tag{6c}$$

$$z_{1,k} + z_{2,k} \leq t_k, \ \frac{(1-\tau_k)L_k}{C_{2,k}} \leq t_k, \quad \forall k \tag{6d}$$

$$t = t_1 = ... = t_K, \ (5b) - (5e). \tag{6e}$$

where $\mathcal{Z} \triangleq \{\mathbf{v}_{RF,k}, \mathbf{U}_{BB}, \mathbf{U}_{RF}, \tau_k, C_{1,k}, t, t_k, z_{1,k}, z_{2,k}\}$.

Moreover, due to the complex expressions with fractional form of the SINR shown in (6c), constraint (6c) is still hard to tackle. To further handle it, we introduce auxiliary variables $\{R_k\}$ and $\{\phi_k\}, \forall k \in \mathcal{K}$. Then constraint (6c) can be equivalently transformed into the following constraints:

$$\frac{\tau_k L_k}{W R_k} \leq z_{2,k}, \quad \forall k \tag{7}$$

$$R_k \leq \log_2(1+\phi_k), \quad \forall k \tag{8}$$

$$\phi_k \leq \frac{P|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_k \mathbf{v}_{RF,k}|^2}{\sum_{k'\neq k} P|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H \mathbf{H}_{k'} \mathbf{v}_{RF,k'}|^2 + \|\mathbf{u}_{BB,k}^H \mathbf{U}_{RF}^H\|^2 \sigma^2}, \ \forall k. \tag{9}$$

By applying the previous reformulation, problem (6) can be equivalently converted to

$$\min_{\tilde{\mathcal{Z}}} \quad t \tag{10a}$$

$$\text{s.t.} \quad (6b), \ (6d), \ (6e), \ (7)-(9), \tag{10b}$$

where $\tilde{\mathcal{Z}} \triangleq \{\mathbf{v}_{RF,k}, \mathbf{U}_{BB}, \mathbf{U}_{RF}, \tau_k, C_{1,k}, t, t_k, z_{1,k}, z_{2,k}, R_k, \phi_k\}$.

To cope with the coupling constraints shown in (10b), we further introduce a set of auxiliary variables $\{\bar{t}_k\}$, $\{\bar{z}_{1,k}\}$, $\{\bar{z}_{2,k}\}$, $\{\bar{\tau}_k\}$, $\{\hat{\tau}_k\}$, $\{\tilde{\tau}_k\}$, $\{\bar{C}_{1,k}\}$, $\{\bar{R}_k\}$, $\{\bar{\phi}_k\}$ and $\{\tilde{\mathbf{u}}_k\}, \forall k \in \mathcal{K}$ which meet the following equality constraints: $\bar{t}_k = t_k$, $\bar{z}_{1,k} = z_{1,k}$, $\bar{z}_{2,k} = z_{2,k}$, $\bar{\tau}_k = \tau_k$, $\hat{\tau}_k = \tau_k$, $\tilde{\tau}_k = \tau_k$, $\bar{C}_{1,k} = C_{1,k}$, $\bar{R}_k = R_k$, $\bar{\phi}_k = \phi_k$ and $\tilde{\mathbf{u}}_k = \mathbf{U}_{RF}\mathbf{u}_{BB,k}$. We also introduce auxiliary variables $\{\tilde{v}_{k,k'}\}, \forall k, k' \in \mathcal{K}$ which satisfies $\tilde{v}_{k,k'} = \tilde{\mathbf{u}}_k^H \mathbf{H}_{k'} \mathbf{v}_{RF,k'}$. Then problem (10) can be equivalently expressed as

$$\min_{\chi} \ t \tag{11a}$$

$$\text{s.t.} \quad \frac{\bar{\tau}_k L_k}{\bar{C}_{1,k}} \leq z_{1,k}, \quad \forall k \tag{11b}$$

$$\frac{\tilde{\tau}_k L_k}{W R_k} \leq z_{2,k}, \quad \forall k \tag{11c}$$

$$\bar{z}_{1,k} + \bar{z}_{2,k} \leq t_k, \ \frac{(1-\hat{\tau}_k)L_k}{C_{2,k}} \leq \bar{t}_k, \quad \forall k \tag{11d}$$

$$\bar{R}_k \leq \log_2(1+\phi_k), \ 0 \leq \tau_k \leq 1, \quad \forall k \tag{11e}$$

$$\sum_{k'\neq k} |\tilde{v}_{k,k'}|^2 + \|\tilde{\mathbf{u}}_k\|^2 \frac{\sigma^2}{P} - \frac{|\tilde{v}_{k,k}|^2}{\bar{\phi}_k} \leq 0, \quad \forall k \tag{11f}$$

$$\sum_{k=1}^{K} C_{1,k} \leq C_{\max} \tag{11g}$$

$$|\mathbf{v}_{RF,k}(i)| = 1, \ |\mathbf{U}_{RF}(i,j)| = 1, \ \forall k, i, j \tag{11h}$$

$$t = t_k = \bar{t}_k, \ z_{1,k} = \bar{z}_{1,k}, \ z_{2,k} = \bar{z}_{2,k}, \ \tau_k = \bar{\tau}_k = \hat{\tau}_k = \tilde{\tau}_k, \quad \forall k \tag{11i}$$

$$\bar{C}_{1,k} = C_{1,k}, \ \bar{R}_k = R_k, \ \bar{\phi}_k = \phi_k, \ \tilde{\mathbf{u}}_k = \mathbf{U}_{RF}\mathbf{u}_{BB,k}, \quad \forall k \tag{11j}$$

$$\tilde{v}_{k,k'} = \tilde{\mathbf{u}}_k^H \mathbf{H}_{k'} \mathbf{v}_{RF,k'}, \quad \forall k, k' \tag{11k}$$

where $\chi \triangleq \{\mathbf{v}_{RF,k}, \mathbf{U}_{BB}, \mathbf{U}_{RF}, C_{1,k}, \bar{C}_{1,k}, t, t_k, \bar{t}_k, z_{1,k}, \bar{z}_{1,k}, z_{2,k}, \bar{z}_{2,k}, R_k, \bar{R}_k, \phi_k, \bar{\phi}_k, \tau_k, \bar{\tau}_k, \hat{\tau}_k, \tilde{\tau}_k, \tilde{\mathbf{u}}_k, \tilde{v}_{k,k'}\}$.

B. Proposed PDD-based Algorithm

Next, we put forward an efficient joint hybrid beamforming and resource allocation algorithm based on the PDD approach. The framework of the PDD method can be found in Appendix A.

Based on the PDD framework, we move the equality constraints (11i)-(11k) into the objective function together with Lagrange multipliers $\{\lambda_{t_k}\}$, $\{\lambda_{\bar{t}_k}\}$, $\{\lambda_{z_{1,k}}\}$, $\{\lambda_{z_{2,k}}\}$, $\{\lambda_{\bar{\tau}_k}\}$, $\{\lambda_{\hat{\tau}_k}\}$, $\{\lambda_{\tilde{\tau}_k}\}$, $\{\lambda_{C_{1,k}}\}$, $\{\lambda_{R_k}\}$, $\{\lambda_{\phi_k}\}$, $\{\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k}\}$, $\{\lambda_{\tilde{v}_{k,k'}}\}$ and the penalty coefficient ρ. Then the resulting AL problem can be expressed as

$$\min_{\chi} \; t + P_\rho(\chi) \tag{12a}$$

$$\text{s.t.} \quad (11b) - (11h) \tag{12b}$$

where $P_\rho(\chi) \triangleq \frac{1}{2\rho}\sum_{k=1}^{K}(|t - t_k + \rho\lambda_{t_k}|^2 + |t - \bar{t}_k + \rho\lambda_{\bar{t}_k}|^2 + |z_{1,k} - \bar{z}_{1,k} + \rho\lambda_{z_{1,k}}|^2 + |z_{2,k} - \bar{z}_{2,k} + \rho\lambda_{z_{2,k}}|^2 + |\tau_k - \bar{\tau}_k + \rho\lambda_{\bar{\tau}_k}|^2 + |\tau_k - \hat{\tau}_k + \rho\lambda_{\hat{\tau}_k}|^2 + |\tau_k - \tilde{\tau}_k + \rho\lambda_{\tilde{\tau}_k}|^2 + |C_{1,k} - \bar{C}_{1,k} + \rho\lambda_{C_{1,k}}|^2 + |R_k - \bar{R}_k + \rho\lambda_{R_k}|^2 + |\phi_k - \bar{\phi}_k + \rho\lambda_{\phi_k}|^2 + \|\tilde{\mathbf{u}}_k - \mathbf{U}_{RF}\mathbf{u}_{BB,k} + \rho\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k}\|^2) + \frac{1}{2\rho}\sum_{k=1}^{K}\sum_{k'=1}^{K}|\tilde{v}_{k,k'} - \tilde{\mathbf{u}}_k^H\mathbf{H}_{k'}\mathbf{v}_{RF,k'} + \rho\lambda_{\tilde{v}_{k,k'}}|^2$.

The main objective of the proposed PDD-based joint design is to efficiently solve the problem (12) with fixed dual variables and penalty parameter in the inner loop.

C. Proposed CCCP-based Algorithm for Solving Problem (12) in the inner loop

Now we turn attention to deal with problem (12). This problem is challenging due to the nonconvex constraints in (11b), (11c) and (11f). Note that these nonconvex constraints can be seen as difference-of-convex (DC) functions. By using the linearization operation, these constraints can be approximated as convex ones, then we can solve the AL problem based on CCCP-based iterative algorithm. First, let us focus on constraint (11b), where it can be rewritten as

$$f_1(\mathbf{d}_k) - f_2(\mathbf{d}_k) \le 0, \quad \forall k \tag{13}$$

where

$$f_1(\mathbf{d}_k) \triangleq \bar{\tau}_k + \frac{\left(\frac{\bar{C}_{1,k}}{L_k} - z_{1,k}\right)^2}{4}, \tag{14a}$$

$$f_2(\mathbf{d}_k) \triangleq \frac{\left(\frac{\bar{C}_{1,k}}{L_k} + z_{1,k}\right)^2}{4}, \tag{14b}$$

and $\mathbf{d}_k = [\bar{\tau}_k, \bar{C}_{1,k}, z_{1,k}]^T$. We approximate the convex function $f_2(\mathbf{d}_k)$ in the lth iteration by its first order Taylor expansion around the current point $\mathbf{d}_k^l = [\bar{\tau}_k, \bar{C}_{1,k}^l, z_{1,k}^l]^T$, denoted as

$$\begin{aligned}\hat{f}_2(\mathbf{d}_k, \mathbf{d}_k^l) &\triangleq f_2(\mathbf{d}_k^l) + \nabla f_2^T(\mathbf{d}_k^l)(\mathbf{d}_k - \mathbf{d}_k^l) \\ &= \frac{\left(\frac{\bar{C}_{1,k}}{L_k} + z_{1,k}\right)\left(\frac{\bar{C}_{1,k}^l}{L_k} + z_{1,k}^l\right)}{2} - \frac{\left(\frac{\bar{C}_{1,k}^l}{L_k} + z_{1,k}^l\right)^2}{4}.\end{aligned} \tag{15}$$

Based on the above results, constraint (11b) can be approximated as a convex constraint as

$$f_1(\mathbf{d}_k) - \hat{f}_2(\mathbf{d}_k, \mathbf{d}_k^l) \le 0, \quad \forall k. \tag{16}$$

By following the same approach, constraints (11c) and (11f) can be approximated as (17) and (18), respectively.

Based on the concept of CCCP [33], [34], problem (12) in the lth iteration can be expressed as the following approximated convex one

$$\min_{\chi} \; t + P_\rho(\chi) \tag{19a}$$

$$\text{s.t.} \quad (11d) - (11e), (11g) - (11h), (16), (17), (18). \tag{19b}$$

Then, in each iteration of the proposed CCCP-based algorithm, we partition the design variables into three blocks, and these block of variables are updated in a BCD fashion [36] in order to minimize the objective function. Details of the derivation of BCD iterations in the CCCP-based algorithm are shown in Appendix B. In **Algorithm 1**, we summarize the CCCP-based algorithm which is adopted in the inner loop of the proposed PDD-based algorithm.

Algorithm 1 Proposed CCCP-based algorithm for problem (19)

1. Define the tolerance of accuracy ϵ_1. Initialize the algorithm with a feasible point.
2. **repeat**
3. – Update $\{\bar{z}_{1,k}, \bar{z}_{2,k}, t_k\}$, $\{\bar{R}_k, \phi_k\}$, $\{\mathbf{u}_{BB,k}\}$, $\{\mathbf{v}_{RF,k}\}$, $\{\tau_k\}$, and $\{C_{1,k}\}$ in **Step 1**.
4. – Update $\{\bar{\tau}_k, \bar{C}_{1,k}, z_{1,k}\}$, $\{\tilde{\tau}_k, R_k, z_{2,k}\}$, $\{\hat{\tau}_k, \bar{t}_k\}$, and $\{\tilde{v}_{k,k'}, \tilde{\mathbf{u}}_k, \bar{\phi}_k\}$ in **Step 2**.
5. – Update t and $\mathbf{U}_{RF}$ in **Step 3**.
6. **Until** the difference successive values of the objective function is less than ϵ_1.

D. Summary of the Proposed PDD-based Algorithm

Let us define the constraint violation $\|\mathbf{h}(\mathbf{x}^m)\|_\infty$ as (20). After running **Algorithm 1** in the inner loop of the proposed PDD-based algorithm, the penalty parameter ρ is updated according to $\rho^{m+1} = c\rho^m$ $(0 < c < 1)$ based on the constraint violation condition and the dual variables are updated according to

$$\begin{aligned}
&\lambda_{t_k}^{m+1} = \lambda_{t_k}^m + \frac{1}{\rho^m}(t - t_k), \; \lambda_{\bar{t}_k}^{m+1} = \lambda_{\bar{t}_k}^m + \frac{1}{\rho^m}(t - \bar{t}_k), \\
&\lambda_{z_{1,k}}^{m+1} = \lambda_{z_{1,k}}^m + \frac{1}{\rho^m}(z_{1,k} - \bar{z}_{1,k}), \\
&\lambda_{z_{2,k}}^{m+1} = \lambda_{z_{2,k}}^m + \frac{1}{\rho^m}(z_{2,k} - \bar{z}_{2,k}), \\
&\lambda_{\bar{\tau}_k}^{m+1} = \lambda_{\bar{\tau}_k}^m + \frac{1}{\rho^m}(\tau_k - \bar{\tau}_k), \; \lambda_{\hat{\tau}_k}^{m+1} = \lambda_{\hat{\tau}_k}^m + \frac{1}{\rho^m}(\tau_k - \hat{\tau}_k), \\
&\lambda_{\tilde{\tau}_k}^{m+1} = \lambda_{\tilde{\tau}_k}^m + \frac{1}{\rho^m}(\tau_k - \tilde{\tau}_k), \\
&\lambda_{C_{1,k}}^{m+1} = \lambda_{C_{1,k}}^m + \frac{1}{\rho^m}(C_{1,k} - \bar{C}_{1,k}), \\
&\lambda_{R_k}^{m+1} = \lambda_{R_k}^m + \frac{1}{\rho^m}(R_k - \bar{R}_k), \; \lambda_{\phi_k}^{m+1} = \lambda_{\phi_k}^m + \frac{1}{\rho^m}(\phi_k - \bar{\phi}_k), \\
&\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k}^{m+1} = \boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k}^m + \frac{1}{\rho^m}(\tilde{\mathbf{u}}_k - \mathbf{U}_{RF}\mathbf{U}_{BB,k}), \\
&\lambda_{\tilde{v}_{k,k'}}^{m+1} = \lambda_{\tilde{v}_{k,k'}}^m + \frac{1}{\rho^m}(\tilde{v}_{k,k'} - \tilde{\mathbf{u}}_k^H\mathbf{H}_{k'}\mathbf{v}_{RF,k'})
\end{aligned} \tag{21}$$

$$\frac{L_k}{W}\tilde{\tau}_k + \frac{(R_k - z_{2,k})^2}{4} - \left(\frac{(R_k + z_{2,k})\left(R_k^l + z_{2,k}^l\right)}{2} - \frac{\left(R_k^l + z_{2,k}^l\right)^2}{4} \right) \leq 0, \quad \forall k. \tag{17}$$

$$\sum_{k' \neq k} |\tilde{v}_{k,k'}|^2 + \|\tilde{\mathbf{u}}_k\|^2 \frac{\sigma^2}{P} - \left(\frac{\tilde{v}_{k,k}^{l*}\tilde{v}_{k,k}}{\bar{\phi}_k^l} + \frac{\tilde{v}_{k,k}^{*}\tilde{v}_{k,k}^{l}}{\bar{\phi}_k^l} - \frac{|\tilde{v}_{k,k}^{l}|^2\bar{\phi}_k}{\bar{\phi}_k^{l2}} \right) \leq 0, \quad \forall k. \tag{18}$$

$$\begin{aligned} \|\mathbf{h}(\mathbf{x}^m)\|_\infty = \max\{&|t - t_k|, |t - \bar{t}_k|, |z_{1,k} - \bar{z}_{1,k}|, |z_{2,k} - \bar{z}_{2,k}|, |\tau_k - \bar{\tau}_k|, |\tau_k - \hat{\tau}_k|, |\tau_k - \tilde{\tau}_k|, \\ &|C_{1,k} - \bar{C}_{1,k}|, |R_k - \bar{R}_k|, |\phi_k - \bar{\phi}_k|, \|\tilde{\mathbf{u}}_k - \mathbf{U}_{RF}\mathbf{U}_{BB,k}\|, |\tilde{v}_{k,k'} - \tilde{\mathbf{u}}_k^H \mathbf{H}_{k'}\mathbf{v}_{RF,k'}|\}, \ \forall k, k'. \end{aligned} \tag{20}$$

where m is the number of outer iterations. For the detailed update procedure, please also refer to **Algorithm 2** shown in Appendix A. Based on the above discussions, the flow of the proposed PDD-based algorithm is summarized in Fig. 2.

From the convergence analysis provided in [27], we conclude that the proposed PDD-based joint hybrid beamforming and resource allocation algorithm converges to the set of stationary solutions of problem (12). Here, we omit the detailed proof for simplicity.

IV. Practical Implementation and Complexity analysis

In this section, we elaborate on how the proposed PDD-based algorithm is implemented in practice and analyze its computational complexity. Observe from Fig. 2 that the proposed algorithm contains the update of design variables, dual variables and penalty parameter, and it can be performed by the BS as well as the users in a distributed fashion.

In the initial phase, the BS collects the channel state information and generates a set of feasible variables then disseminates them to all the users. In the inner loop of the PDD-based algorithm, i.e., the proposed CCCP-based algorithm in **Algorithm 1**, the design variables are updated in a distributed fashion. In **Step 1**, each user $k \in \mathcal{K}$ updates $\{\bar{z}_{1,k}, \bar{z}_{2,k}, t_k\}$, $\{\bar{R}_k, \phi_k\}$, $\{\mathbf{u}_{BB,k}\}$, $\{\mathbf{v}_{RF,k}\}$, $\{\tau_k\}$, and $\{C_{1,k}\}$ locally in parallel. In **Step 2**, each user updates $\{\bar{\tau}_k, \bar{C}_{1,k}, z_{1,k}\}$, $\{\tilde{\tau}_k, R_k, z_{2,k}\}$, $\{\hat{\tau}_k, \bar{t}_k\}$, and $\{\tilde{v}_{k,k'}, \tilde{\mathbf{u}}_k, \bar{\phi}_k\}$ individually without the need of information exchange among other users. In **Step 3**, all the users first send the updated design variables to the BS, then the BS optimizes t and $\mathbf{U}_{RF}$ in parallel and broadcasts the results back to the users. After the iteration of the inner loop, the BS and the users update the penalty parameter ρ and the dual variables, respectively, and then exchange these updating results. Based on the above description of the information exchange mechanism, the proposed PDD-based algorithm can be implemented in a distributed way. In particular, the design variables in these three blocks are updated sequentially in a parallel manner for the BS and users.

To analyze the computational complexity of the proposed PDD-based algorithm, we mainly focus our attention on the updating of $\{\tilde{v}_{k,k'}, \tilde{\mathbf{u}}_k, \bar{\phi}_k\}$ and $\mathbf{U}_{RF}$, which dominates the complexity. The complexity of updating $\{\tilde{v}_{k,k'}, \tilde{\mathbf{u}}_k, \bar{\phi}_k\}$ depends on the bisection method used to search the Lagrangian parameter. The number of iterations is $\log_2(\frac{\theta_{0,s}}{\theta_s})$, where $\theta_{0,s}$ is the initial interval size and θ_s denotes the tolerance. Thus, we can conclude that the computational cost for solving this subproblem is roughly $\mathcal{O}\left(\log_2(\frac{\theta_{0,s}}{\theta_s})\right)$. When updating $\mathbf{U}_{RF}$, the proposed one-iteration BCD type algorithm has a complexity of $\mathcal{O}(N_{\mathrm{BS}}^2 N_{\mathrm{s}}^2)$. Therefore, the overall complexity of the proposed algorithm can be expressed as

$$\mathcal{O}\left(I_1 I_2 \left(\log_2\left(\frac{\theta_{0,s}}{\theta_s}\right) + N_{\mathrm{BS}}^2 N_{\mathrm{s}}^2\right)\right), \tag{22}$$

where I_1 and I_2 denote the maximum number of iterations for the inner and outer loops.

V. Simulation results

In this section, the performance of the proposed PDD-based joint hybrid beamforming and resource allocation algorithm is evaluated by means of computer simulations[1]. In the experiments, we consider a mmWave based MEC system supporting multiple mobile users, and a 28GHz mmWave outdoor[2] cellular propagation statistical channel model is adopted. Considering the small-scale time-varying channel matrix, we apply a standard uniform linear antenna array mmWave channel model. Due to the sparsity and high free-space path loss, an extended Saleh-Valenzuela geometric model [38] is adopted in this paper, and the channel matrix $\mathbf{H}_s$ can be expressed as

$$\mathbf{H}_s = \sqrt{\frac{N_t N_r}{L_p}} \sum_{l=1}^{L_p} \alpha_l \mathbf{a}_r(\phi_l^r)\mathbf{a}_t(\phi_l^t)^H, \tag{23}$$

where L_p is the number of distinguishable paths, $\alpha_l \sim \mathcal{CN}(0,1)$ is the complex gain of the l-th path, $\mathbf{a}_r(\phi_l^r)$ and $\mathbf{a}_t(\phi_l^t)$ are the receive and transmit antenna array response vectors, where $\phi_l^r \in [0, 2\pi)$ and $\phi_l^t \in [0, 2\pi)$ are the azimuth angles of arrival and departure (AoAs and AoDs), respectively. Thus, the response vector takes the form

$$\mathbf{a}(\theta) = \frac{1}{N}\left[1, e^{jkd_a \sin(\theta)}, ..., e^{jkd_a(N-1)\sin(\theta)}\right]^T, \tag{24}$$

where $k = 2\pi/\lambda$, λ is the wavelength, and d_a is the antenna spacing. As for the large-scale fading, we adopt a statistical

[1]Note that the proposed algorithm is not necessarily intended to suitable for a specific channel model, but rather to an operating band of frequency (i.e. 30-300GHz) considered for 5G systems, where due to technological limitations, the use of hybrid beamforming structures is essential.

[2]Indoor environment is also suitable in our model, but with different channel parameters [37].

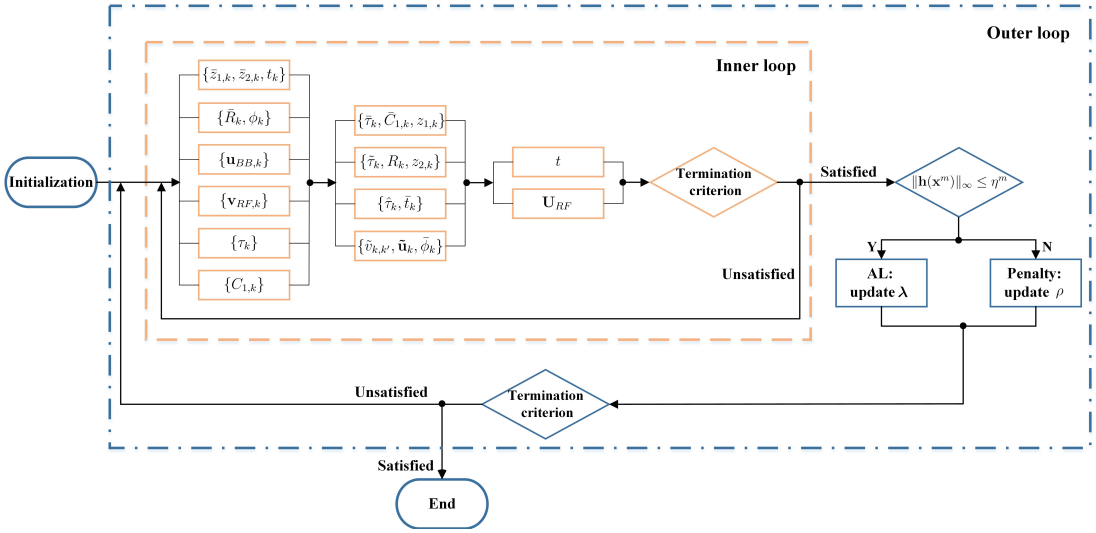

Fig. 2: The flow of the proposed PDD-based algorithm

model, which is detailed in [15]. For each path, the relationship between the omnidirectional path loss PL and the lognormal shadowing ξ is given by

$$PL(d)[\text{dB}] = \alpha + \beta 10\log_{10}(d) + \xi, \tag{25}$$

where d denotes the distance in meters, $\xi \sim \mathcal{N}(0, \sigma_s^2)$ is the lognormal shadowing variance. Each mmWave channel is in a LOS or a NLOS state. The probability functions for these two states can be expressed as

$$p_{\text{LOS}}(d) = e^{-a_{\text{los}}d} \tag{26a}$$

$$p_{\text{NLOS}}(d) = 1 - p_{\text{LOS}}(d) \tag{26b}$$

where a_{los} is a parameter fitting from the real environment [15]. We have $\alpha = 72.0, \beta = 2.92$, and $\sigma_s = 8.7$ dB in the NLOS state, and $\alpha = 61.4, \beta = 2$, and $\sigma_s = 5.8$ dB in the LOS state.

In the simulations, we set $N_{\text{BS}} = 32, N_{\text{s}} = 8$, and $N_{\text{user}} = 4$. We also set $W = 50$MHz and $L_p = 20$. For simplicity, the total size of the input computational task is set to be 1×10^7 bits for all the users. In addition, the computational resource of the MEC server is given by 1×10^8 bits/s, and the local capability is 0.4×10^7 bits/s for all the users [39], [40]. The transmit power P is set to be 0.25mW, and the noise power level is $\sigma^2 = 10^{-11}$W. For the PDD-based algorithm, the tolerance parameters are chosen as $\epsilon_1 = 1 \times 10^{-3}$. The initial penalty parameters are set to be $\rho^0 = 2$ with $c = 0.6$. In addition, we also set $\eta^0 = 0.1$ and $\eta^{m+1} = 0.7\eta^m$. We also develop two benchmarks for comparison.

1) The CM-based algorithm, which is a heuristic algorithm. The details of this algorithm are shown in Appendix C.
2) The TDMA-based algorithm, which can be described as follows: All the users share the same frequency channel by dividing the task processing into different time slots. In each time slot, one portion of tasks is processed locally at the user side while the rest is offloaded to the BS and processed at the MEC server utilising its full capability. During each time slot, the system model degrades as a simplified single-user mmWave MEC model, and the delay minimization problem in the single user case is well investigated in our paper [41] by using an iterative weighted mean-square error minimization (WMMSE) approach. When all the users' tasks have been processed, the system delay can be calculated by adding all the task's processing time.

Let us first examine the convergence of the proposed PDD-based algorithm. We assume supporting four users by the system, and all the users are 40 meters away from the BS. In Fig. 3(a), we show the value of the objective function (19a) versus the number of outer iterations for the proposed algorithm, which can be seen to converge within 20 iterations. Fig. 3(b) shows the corresponding value of the constraint violation indicator. It shows that after 40 iterations, the penalty terms decrease to a value below 10^{-5}, which demonstrates that the proposed PDD-based algorithm can tackle the equality constraints efficiently. Based on the results, we conclude that our proposed algorithm has a rapid convergence. Next, we

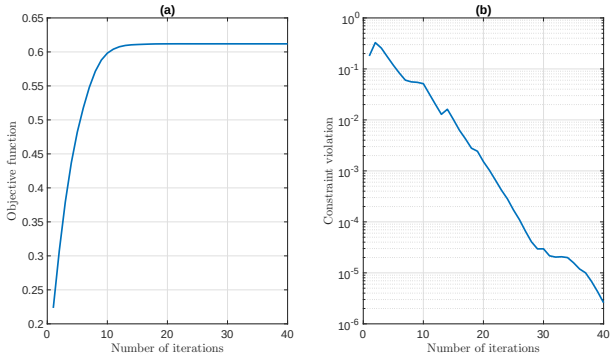

Fig. 3: (a) Objective function value and (b) constraint violation versus the number of outer iterations for the proposed PDD-based algorithm.

compare the proposed PDD-based joint design and the heuristic algorithm in terms of the maximum delay performance. We first consider a two user case, where the distance between user 1 and the BS is fixed to be 10 meters, while the distance between user 2 and the BS varies from 20 to 80 meters. Fig.

4 shows that when the distance between user 2 and the BS increases, the maximum delay becomes larger for the proposed PDD-based joint design, the CM-based heuristic algorithm and the TDMA-based algorithm. This is due to the fact that the channel condition of $\mathbf{H}_2$ becomes poor with the increasing of the distance, and it results in a longer transmission delay for user 2. As the distance between user 2 and the BS increases, the gap between the proposed joint design and the CM-based heuristic algorithm increases. We also notice that the TDMA-baed algorithm performs well in this scenario, since the number of users is small, and the distance between user 1 and the BS is short, resulting a short processing delay.

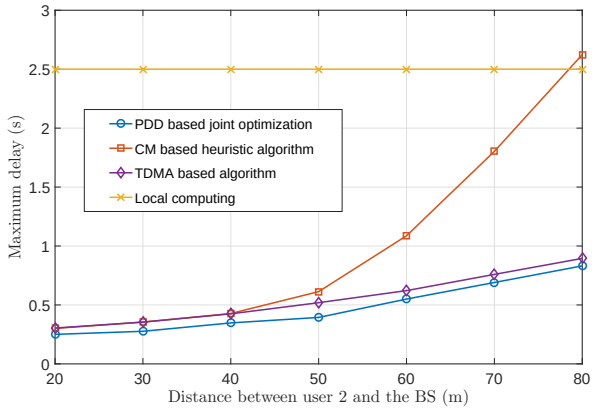

Fig. 4: The maximum system delay versus the location of user 2

Fig. 5 illustrates the optimal offloading ratio and resource allocation (the ratio of computational resource at the MEC server allocated to user 2) versus the distance between user 2 and the BS. It can be observed in Fig. 5(a) that the optimal offloading ratio of the task of user 2 decreases with the increase of the distance. This is rather intuitive due to the fact that more computational task will be processed at the user locally when the channel condition gets worse. Furthermore, it is shown in Fig. 5(b) that the optimal computational resource allocated to user 2 increases with the increase of the distance. The reason for this outcome is that more computational resources are needed to compensate the transmission delay in mmWave channels. Moreover, there exists a tradeoff between the offloading ratio and resource allocation.

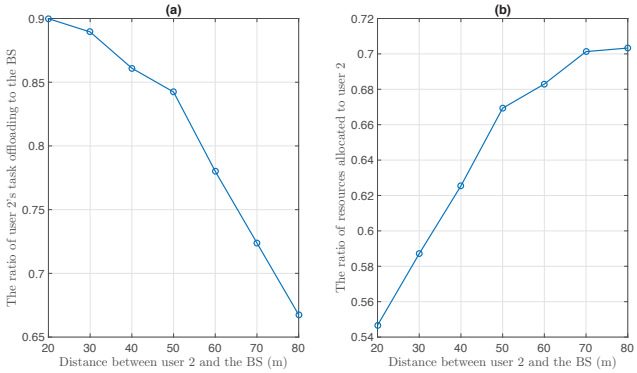

Fig. 5: Offloading ratio and resource allocation

We then study the impact of the number of users. In Fig. 6, we plot the system's maximum delay versus the number of users. All the users cover an area of circle with radius of 40m/60m. The proposed PDD-based algorithm and the compared algorithms are implemented. Firstly, except for the local computing, the maximum delay for all algorithms increases as the number of users due to the computational budget, while the maximum delay of the local computing is invariant. Secondly, by comparing the PDD-based algorithm with the CM-based and TDMA-based algorithm, we can observe that the performance gain becomes more evident upon increasing number of users, which demonstrates the benefits of the proposed joint design. Thirdly, for a large number of users, the proposed CM-based and TDMA-based algorithm perform even worse than the local computing method at a distance of 60m. Fourthly, the TDMA-based algorithm shows similar performance to the CM-based algorithm in a short distance, and performs better in a large distance. This phenomenon is joint determined by the channel condition, resource allocation strategy and the capability of the MEC server. Finally, the proposed PDD-based algorithm provides the best performance among the algorithms analyzed even in the face of poor channel conditions and a large number of users due to the associated joint design.

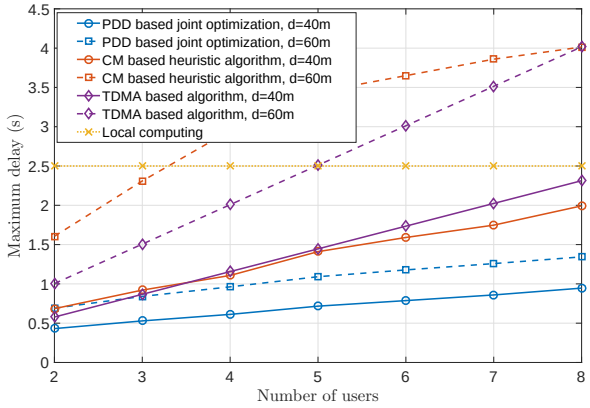

Fig. 6: The maximum system delay versus the number of users

In Fig. 7, we plot the system's maximum delay versus the computational capability of mobile users. In this simulation, we assume that there are four mobile users in the system, while the computational capability varies in the range between 1×10^6 bits/s and 6×10^6 bits/s. From the results, we can observe that the system delay of all the four algorithms decreases since the local processing time decreases. Furthermore, in contrast to the local computing method, the proposed PDD-based algorithm is not sensitive to the user's computational capability due to the powerful computational capability at the MEC server.

Fig. 8 presents the system's maximum delay versus the computational capability of MEC. We assume that there are four mobile users in the system, while the MEC server computational capability varies in the range between 0.4×10^8 bits/s to 2.4×10^8 bits/s. It can be observed that the system delay of both the proposed PDD-based joint design and the TDMA-based algorithm decreases with the computational capability

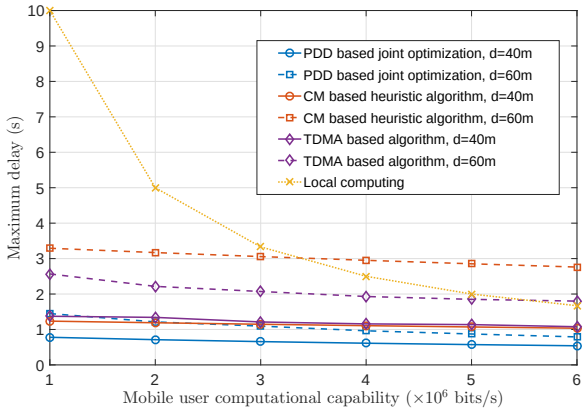

Fig. 7: The maximum system delay versus the computational capability of users

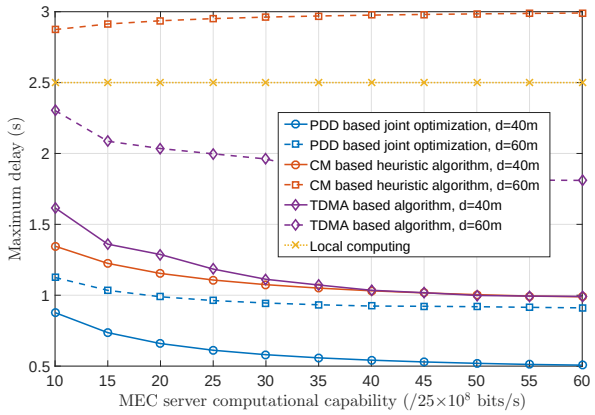

Fig. 8: The maximum system delay versus the computational capability of MEC

of the MEC server, since more resources are available at the MEC server. For the CM-based algorithm, the system delay decreases with the computational capability of the MEC server when $d = 40$m, but increases a little at a distance of 60m, this is due to the fact that the offloading ratio will increase with the MEC server computaional capability, causing large transimission latency in a bad channel condition. Additionally, when the computational capability of MEC becomes sufficiently high, the curves for the PDD-based algorithm tend towards a fixed value, i.e. the system's delay becomes limited by the radio resources, which reveals that once the system performance is constrained by the radio resources, having redundant MEC computational resources becomes unnecessary.

VI. Conclusion

An efficent joint beamforming and resource allocation design has been conceived for multi-user mmWave based MEC systems. We reformulated the system delay minimization problem by invoking a series of suitable transformations. Then, based on the PDD technique, we developed an innovative distributed algorithm for the analog beamforming vectors at the users, the analog and digital beamforming matrices at the BS, the task offloading ratios, and the resource allocation at the MEC server. Our simulation results indicate the superiority of the joint design, and demonstrate the benefits of combining MEC with mmWave communications.

Appendix A
Framework of the PDD method

The PDD method is a double-loop iterative algorithm which can address nonconvex nonsmooth problems with nonconvex coupling constraints. Let us consider the following optimization problem: the minimization of a nonconvex objective function $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$ subject to equality constraint $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ and possibly nonconvex inequality constraints $\mathbf{g}_i(\mathbf{x}_i) \leq \mathbf{0}$, i.e.,

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{h}(\mathbf{x}) = \mathbf{0}, \\ & \mathbf{g}_i(\mathbf{x}_i) \leq \mathbf{0},\ i = 1, 2, \ldots, n. \end{aligned} \tag{27}$$

For solving problem (27), the detailed steps of the PDD method are presented in **Algorithm 2**, where the inner loop of this algorithm (Step 3) addresses the AL optimization problem in a BCD fashion. This is the key to implement the PDD method, which invokes an iterative optimization algorithm to solve the following problem to some accuracy ϵ^m

$$\begin{aligned} & P(\rho^m, \boldsymbol{\lambda}^m) \\ & \min_{\mathbf{x}\in\mathcal{X}} \quad \{\mathcal{L}_m(\mathbf{x}) \triangleq f(\mathbf{x}) + \boldsymbol{\lambda}^{mT}\mathbf{h}(\mathbf{x}) + \frac{1}{2\rho^m}\|\mathbf{h}(\mathbf{x})\|^2\}, \end{aligned} \tag{28}$$

where $\mathcal{L}_m(\mathbf{x})$ is the AL function with dual variable $\boldsymbol{\lambda}^m$ and penalty parameter ρ^m. The outer loop focuses on updating the dual variables or penalty parameter in terms of the constraint violation, i.e., the term $\|\mathbf{h}(\mathbf{x}^m)\|_\infty$[3].

Furthermore, it can be shown that the sequence generated by the PDD method converges to a KKT (stationary) point of problem (27) under suitable constraints condition. The details about the convergence analysis can be found in [27].

Algorithm 2 PDD method for problem (27)

1. **initialize** $\mathbf{x}^0, \rho^0 > 0, \boldsymbol{\lambda}^0$, and set $0 < c < 1$, $m = 1$.
2. **repeat**
3. $\quad \mathbf{x}^m = \text{optimize}(P(\rho^m, \boldsymbol{\lambda}^m), \mathbf{x}^{m-1}, \epsilon^m)$
4. $\quad$ **if** $\|\mathbf{h}(\mathbf{x}^m)\|_\infty \leq \eta^m$
5. $\quad\quad \boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m + \frac{1}{\rho^m}\mathbf{h}(\mathbf{x}^m)$
6. $\quad\quad \rho^{m+1} = \rho^m$
7. $\quad$ **else**
8. $\quad\quad \boldsymbol{\lambda}^{m+1} = \boldsymbol{\lambda}^m$
9. $\quad\quad \rho^{m+1} = c\rho^m$
10. $\quad$ **end**
11. $\quad m = m + 1$
12. **until** some termination criterion is met.

Appendix B
Derivation of updating steps in the inner loop algorithm

In this part, we derive each of the update procedure in the inner loop algorithm.

[3]In practice, it is reasonable to terminate the algorithm based on the progress of the objective value, i.e., $\frac{|\mathcal{L}_m(\mathbf{x}^m) - \mathcal{L}_{m-1}(\mathbf{x}^{m-1})|}{|\mathcal{L}_{m-1}(\mathbf{x}^{m-1})|} \leq \epsilon_m$, or simple by setting the maximum number of iterations.

A. Variables Updating in Step 1

In **Step 1**, we update $\{\bar{z}_{1,k}, \bar{z}_{2,k}, t_k\}$, $\{\bar{R}_k, \phi_k\}$, $\{C_{1,k}\}$, $\{\mathbf{v}_{RF,k}\}$, $\{\mathbf{u}_{BB,k}\}$, and $\{\tau_k\}$ in parallel while fixing the other block of variables. We obtain six independent subproblems. Particularly, the variable blocks $\{\bar{z}_{1,k}, \bar{z}_{2,k}, t_k\}$, $\{\bar{R}_k, \phi_k\}$ and $\{C_{1,k}\}$ can be solved separately based on the Lagrange multiplier method, $\{\mathbf{v}_{RF,k}\}$ can be solved based on the one-iteration BCD-type algorithm [42], while $\{\mathbf{u}_{BB,k}\}$ and $\{\tau_k\}$ can be solved in closed-form solution by taking advantage of the first-order optimality condition. Detailed derivations are demonstrated as follows:

1) **Subproblem for** $\{\bar{z}_{1,k}, \bar{z}_{2,k}, t_k\}$: The subproblem regarding to $\{\bar{z}_{1,k}, \bar{z}_{2,k}, t_k\}$ can be decomposed over each user. As a result, for each $k \in \mathcal{K}$, we have the following optimization problem

$$\min_{\bar{z}_{1,k}, \bar{z}_{2,k}, t_k} |z_{1,k} - \bar{z}_{1,k} + \rho\lambda_{z_{1,k}}|^2 + |z_{2,k} - \bar{z}_{2,k} + \rho\lambda_{z_{2,k}}|^2 + |t - t_k + \rho\lambda_{t_k}|^2 \tag{29a}$$

$$\text{s.t.} \quad \bar{z}_{1,k} + \bar{z}_{2,k} \leq t_k. \tag{29b}$$

Note that there exists only one constraint in this subproblem, so it can be solved in closed form through the Lagrange multiplier method. The corresponding Lagrange function can be written as

$$L(\bar{z}_{1,k}, \bar{z}_{2,k}, t_k, \lambda_{1,k}) \triangleq |z_{1,k} - \bar{z}_{1,k} + \rho\lambda_{z_{1,k}}|^2 + |z_{2,k} - \bar{z}_{2,k} + \rho\lambda_{z_{2,k}}|^2 + |t - t_k + \rho\lambda_{t_k}|^2 + \lambda_{1,k}(\bar{z}_{1,k} + \bar{z}_{2,k} - t_k), \tag{30}$$

where $\lambda_{1,k} \geq 0$ denotes the Lagrange multiplier for constraint (29b). By examining the first order optimality condition of $L(\bar{z}_{1,k}, \bar{z}_{2,k}, t_k, \lambda_{1,k})$, the optimal value of $\bar{z}_{1,k}, \bar{z}_{2,k}, t_k$ can be derived as

$$\bar{z}_{1,k}(\lambda_{1,k}) = \frac{2(z_{1,k} + \rho\lambda_{z_{1,k}}) - \lambda_{1,k}}{2}, \tag{31}$$

$$\bar{z}_{2,k}(\lambda_{1,k}) = \frac{2(z_{2,k} + \rho\lambda_{z_{2,k}}) - \lambda_{1,k}}{2}, \tag{32}$$

$$t_k(\lambda_{1,k}) = \frac{2(t + \rho\lambda_{t_k}) + \lambda_{1,k}}{2}. \tag{33}$$

Let us denote the optimal $\lambda_{1,k}$ as $\lambda^\star_{1,k}$, then $\lambda^\star_{1,k}$ is determined to fulfill the complementary slackness condition of (29b), and can be obtained as (34).

2) **Subproblem for** $\{\bar{R}_k, \phi_k\}$: The subproblem with respect to $\{\bar{R}_k, \phi_k\}, \forall k \in \mathcal{K}$ is a convex problem and it is given by

$$\min_{\bar{R}_k, \phi_k} |R_k - \bar{R}_k + \rho\lambda_{R_k}|^2 + |\phi_k - \bar{\phi}_k + \rho\lambda_{\phi_k}|^2 \tag{35a}$$

$$\text{s.t.} \quad \bar{R}_k \leq \log_2(1 + \phi_k). \tag{35b}$$

Attaching a Lagrange multiplier $\lambda_{2,k} \geq 0$ to constraint (35b), we get the optimal $\bar{R}_k$ and ϕ_k as (36) and (37). In this case, the optimal $\lambda_{2,k}$ can be obtained easily by using the bisection procedure.

3) **Subproblem for** $\{C_{1,k}\}$: The subproblem regarding to $\{C_{1,k}\}, \forall k \in \mathcal{K}$ is formulated as

$$\min_{C_{1,k}} \sum_{k=1}^{K} |C_{1,k} - \bar{C}_{1,k} + \rho\lambda_{C_{1,k}}|^2 \tag{38a}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} C_{1,k} \leq C_{\max}. \tag{38b}$$

By introducing a Lagrange multiplier $\lambda_{3,k} \geq 0$ to constraint (38b), and the optimal $C_{1,k}$ and $\lambda_{3,k}$ can be derived as

$$C_{1,k}(\lambda_{3,k}) = \frac{2(\bar{C}_{1,k} - \rho\lambda_{C_{1,k}}) - \lambda_{3,k}}{2}, \tag{39}$$

$$\lambda_{3,k} = \max\left\{0, \frac{\sum_{k=1}^{K} 2(\bar{C}_{1,k} - \rho\lambda_{C_{1,k}}) - 2C_{\max}}{K}\right\}. \tag{40}$$

4) **Subproblem for** $\{\mathbf{u}_{BB,k}\}$: The corresponding subproblem for $\mathbf{u}_{BB,k}$ is decoupled over each user, and then it is given by

$$\min_{\mathbf{u}_{BB,k}} \|\tilde{\mathbf{u}}_k - \mathbf{U}_{RF}\mathbf{u}_{BB,k} + \rho\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k}\|^2. \tag{41}$$

Note that the above problem is an unconstrained quadratic optimization problem, we can compute the optimal $\mathbf{u}_{BB,k}$ as

$$\mathbf{u}_{BB,k} = (\mathbf{U}^H_{RF}\mathbf{U}_{RF})^{-1}\mathbf{U}^H_{RF}(\tilde{\mathbf{u}}_k + \rho\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k}). \tag{42}$$

5) **Subproblem for** $\{\mathbf{v}_{RF,k}\}$: The corresponding subproblem for this variable set is decoupled over each user, and then for each $k \in \mathcal{K}$, the subproblem is given by

$$\min_{\mathbf{v}_{RF,k}} \sum_{k'=1}^{K} |\tilde{v}_{k',k} - \tilde{\mathbf{u}}^H_{k'}\mathbf{H}_k\mathbf{v}_{RF,k} + \rho\lambda_{\tilde{v}_{k',k}}|^2 \tag{43a}$$

$$\text{s.t.} \quad |\mathbf{v}_{RF,k}(i)| = 1, \quad \forall i. \tag{43b}$$

This subproblem is a quadratic optimization problem with unit modulus constraints, and is complicated mainly due to the constant modulus constraints, which is highly nonconvex. To address this problem, we rearrange it in a more tractable form as

$$\min_{\mathbf{v}_{RF,k}} \text{Tr}\left(\mathbf{v}^H_{RF,k}\mathbf{C}\mathbf{v}_{RF,k}\mathbf{P}\right) - 2\Re e\left\{\text{Tr}\left(\mathbf{v}^H_{RF,k}\mathbf{Q}\right)\right\} \tag{44a}$$

$$\text{s.t.} \quad |\mathbf{v}_{RF,k}(i)| = 1, \quad \forall i. \tag{44b}$$

where $\mathbf{C} = \sum_{k'=1}^{K} \mathbf{H}^H_k\tilde{\mathbf{u}}_{k'}\tilde{\mathbf{u}}^H_{k'}\mathbf{H}_k$, $\mathbf{P} = \mathbf{I}$, and $\mathbf{Q} = \sum_{k'=1}^{K} \mathbf{H}^H_k\tilde{\mathbf{u}}_{k'}(\tilde{v}_{k',k} + \rho\lambda_{\tilde{v}_{k',k}})$.Then we apply the one-iteration BCD-type algorithm shown in [42, Appendix B] to recursively solve this problem, i.e., at each step we update one effective entry of $\mathbf{v}_{RF}$ while fixing the others.

6) **Subproblem for** τ_k: The subproblem with respect to $\tau_k, \forall k \in \mathcal{K}$ can be expressed as

$$\min_{\tau_k} |\tau_k - \bar{\tau}_k + \rho\lambda_{\bar{\tau}_k}|^2 + |\tau_k - \hat{\tau}_k + \rho\lambda_{\hat{\tau}_k}|^2 + |\tau_k - \tilde{\tau}_k + \rho\lambda_{\tilde{\tau}_k}|^2 \tag{45a}$$

$$\text{s.t.} \quad 0 \leq \tau_k \leq 1. \tag{45b}$$

Define $\pi(\tau_k) \triangleq |\tau_k - \bar{\tau}_k + \rho\lambda_{\bar{\tau}_k}|^2 + |\tau_k - \hat{\tau}_k + \rho\lambda_{\hat{\tau}_k}|^2 + |\tau_k - \tilde{\tau}_k + \rho\lambda_{\tilde{\tau}_k}|^2$, and by employing the first-order optimality condition, the optimal value of τ_k can be derived as (46).

$$\lambda_{1,k}^{\star} = \max\left\{0, \frac{2(z_{1,k}+\rho\lambda_{z_{1,k}}) + 2(z_{2,k}+\rho\lambda_{z_{2,k}}) - 2(t+\rho\lambda_{t_k})}{3}\right\}. \tag{34}$$

$$\bar{R}_k(\lambda_{2,k}) = \frac{2(R_k+\rho\lambda_{R_k}) - \lambda_{2,k}}{2}, \tag{36}$$

$$\phi_k(\lambda_{2,k}) = \frac{\bar{\phi}_k - 1 - \rho\lambda_{\phi_k} + \sqrt{(1+\rho\lambda_{\phi_k}-\bar{\phi}_k)^2 + 4(\bar{\phi}_k + \frac{\lambda_{2,k}}{2\ln 2}) - \rho\lambda_{\phi_k}}}{2}. \tag{37}$$

$$\tau_k = \begin{cases} \frac{\bar{\tau}_k+\hat{\tau}_k+\tilde{\tau}_k-2\rho(\lambda_{\bar{\tau}_k}+\lambda_{\hat{\tau}_k}+\lambda_{\tilde{\tau}_k})}{3} & \text{if } 0 \le \frac{\bar{\tau}_k+\hat{\tau}_k+\tilde{\tau}_k-2\rho(\lambda_{\bar{\tau}_k}+\lambda_{\hat{\tau}_k}+\lambda_{\tilde{\tau}_k})}{3} \le 1 \\ 1 & \text{else if } \pi(1) < \pi(0) \\ 0 & \text{else} \end{cases} \tag{46}$$

B. Variables Updating in Step 2

In **Step 2**, we update $\{\bar{\tau}_k, \bar{C}_{1,k}, z_{1,k}\}$, $\{\tilde{\tau}_k, R_k, z_{2,k}\}$, $\{\hat{\tau}_k, \bar{t}_k\}$, and $\{\tilde{v}_{k,k'}, \tilde{\mathbf{u}}_k, \bar{\phi}_k\}$ in parallel by fixing the other block of variables. In this step, the four subproblems are all solved separately based on the Lagrange multiplier method. Detailed derivations are as follows:

1) **Subproblem for** $\{\bar{\tau}_k, \bar{C}_{1,k}, z_{1,k}\}$: The subproblem for $\{\bar{\tau}_k, \bar{C}_{1,k}, z_{1,k}\}$ can be decomposed over each mobile user. As a result, for each $k \in \mathcal{K}$, we have the following optimization problem (47).
Similarly, this subproblem can be solved with the aid of the Lagrange multiplier method. By introducing the Lagrange multiplier $\lambda_{4,k} \ge 0$ for constraint (47b), and examining the first order optimality condition, the optimal value of $\bar{\tau}_k, \bar{C}_{1,k}, z_{1,k}$ can be derived as (48), (49) and (50).

$$\bar{\tau}_k(\lambda_{4,k}) = \frac{2(\tau_k+\rho\lambda_{\bar{\tau}_k}) - \lambda_{4,k}}{2}, \tag{48}$$

$$z_{1,k}(\lambda_{4,k}) = \frac{4(\bar{z}_{1,k}-\rho\lambda_{z_{1,k}}) + \lambda_{4,k}\left(\frac{\bar{C}_{1,k}}{L_k} + \frac{\bar{C}_{1,k}^l}{L_k} + z_{1,k}^l\right)}{4+\lambda_{4,k}}. \tag{50}$$

The optimal $\lambda_{4,k}$ is determined to fulfill the complementary slackness condition of (47b), and can be obtained via bisection search.

2) **Subproblem for** $\{\tilde{\tau}_k, R_k, z_{2,k}\}$: The subproblem for $\{\tilde{\tau}_k, R_k, z_{2,k}\}$ can be decoupled over each mobile user. Therefore, for each $k \in \mathcal{K}$, we have the following optimization problem (51). By introducing a Lagrange multiplier $\lambda_{5,k} \ge 0$ to constraint (51b), we can derive the optimal $\tilde{\tau}_k, R_k$ and $z_{2,k}$ as (52), (53) and (54).

$$\tilde{\tau}_k(\lambda_{5,k}) = \frac{2(\tau_k+\rho\lambda_{\tilde{\tau}_k}) - \lambda_{5,k}\frac{L_k}{W}}{2}, \tag{52}$$

$$z_{2,k}(\lambda_{5,k}) = \frac{4(\bar{z}_{2,k}-\rho\lambda_{z_{2,k}}) + \lambda_{5,k}(R_k + R_k^l + z_{2,k}^l)}{4+\lambda_{5,k}}, \tag{54}$$

where the optimal $\lambda_{5,k}$ can be obtained easily using the bisection procedure.

3) **Subproblem for** $\{\hat{\tau}_k, \bar{t}_k\}$: This subproblem is decoupled across each user. Then the following problem is solved for each $\{\hat{\tau}_k, \bar{t}_k\}, \forall k \in \mathcal{K}$:

$$\min_{\hat{\tau}_k, \bar{t}_k} \ |\tau_k - \hat{\tau}_k + \rho\lambda_{\hat{\tau}_k}|^2 + |t - \bar{t}_k + \rho\lambda_{\bar{t}_k}|^2 \tag{55a}$$

$$\text{s.t.} \quad \frac{(1-\hat{\tau}_k)L_k}{C_{2,k}} \le \bar{t}_k. \tag{55b}$$

By following the same approach and introducing a Lagrange multiplier $\lambda_{6,k} \ge 0$ to constraint (55b), the optimal $\hat{\tau}_k$ and $\bar{t}_k$ can be derived as

$$\hat{\tau}_k(\lambda_{6,k}) = \frac{2(\tau_k+\rho\lambda_{\hat{\tau}_k}) + \lambda_{6,k}}{2}, \tag{56}$$

$$\bar{t}_k(\lambda_{6,k}) = \frac{2(t+\lambda_{\bar{t}_k}) + \lambda_{6,k}C_{2,k}/L_k}{2}. \tag{57}$$

Due to the complementary slackness condition, the optimal Lagrange multiplier $\lambda_{6,k}^{\star}$ can be expressed in a closed-form, which is given by (58).

4) **Subproblem for** $\{\tilde{v}_{k,k'}, \tilde{\mathbf{u}}_k, \bar{\phi}_k\}$: The corresponding subproblem for this variable set is decoupled over each user, and then for each $k \in \mathcal{K}$, the subproblem is given by (59). By following the same approach and introducing a Lagrange multiplier $\lambda_{7,k} \ge 0$ to constraint (59b), the optimal $\tilde{v}_{k,k'}, \tilde{\mathbf{u}}_k$ and $\bar{\phi}_k$ can be expressed as

$$\bar{\phi}_k(\lambda_{7,k}) = \frac{2\left(\phi_k+\rho\lambda_{\phi_k}\right) - \lambda_{7,k}\frac{|\bar{v}_{k,k}^l|^2}{\bar{\phi}_k^{l2}}}{2}, \tag{60}$$

$$\tilde{\mathbf{u}}_k(\lambda_{7,k}) = \mathbf{A}^{-1}(\lambda_{7,k})\mathbf{b}(\lambda_{7,k}), \tag{61}$$

$$\tilde{v}_{k,k'}(\lambda_{7,k}) = \frac{\tilde{\mathbf{u}}_k^H\mathbf{H}_{k'}\mathbf{v}_{RF,k'} - \rho\lambda_{\tilde{v}_{k,k'}}}{1+\lambda_{7,k}}, \ \forall k' \ne k \tag{62}$$

$$\tilde{v}_{k,k}(\lambda_{7,k}) = \tilde{\mathbf{u}}_k^H\mathbf{H}_k\mathbf{v}_{RF,k} - \rho\lambda_{\tilde{v}_{k,k}} + \lambda_{7,k}\frac{\tilde{v}_{k,k}^l}{\bar{\phi}_k^l}, \tag{63}$$

where $\mathbf{A}(\lambda_{7,k}) \triangleq (1+\lambda_{7,k}\frac{\sigma^2}{P})\mathbf{I} + \lambda_{7,k}\frac{\mathbf{R}_k}{1+\lambda_{7,k}}$, and $\mathbf{b}(\lambda_{7,k}) \triangleq \mathbf{U}_{RF}\mathbf{u}_{BB,k} - \rho\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k} + \lambda_{7,k} \times \frac{\sum_{k'\ne k}\mathbf{H}_{k'}\mathbf{v}_{RF,k'}\rho\lambda_{\tilde{v}_{k,k'}}}{1+\lambda_{7,k}} + \lambda_{7,k}\mathbf{H}_k\mathbf{v}_{RF,k}\frac{\tilde{v}_{k,k}^{l*}}{\bar{\phi}_k^l}$. We also let $\mathbf{R}_k \triangleq \sum_{k'\ne k}\mathbf{H}_{k'}\mathbf{v}_{RF,k'}\mathbf{v}_{RF,k'}^H\mathbf{H}_{k'}^H$. The optimal multiplier $\lambda_{7,k}$ should be determined such that the comple-

$$\min_{\bar{\tau}_k,\bar{C}_{1,k},z_{1,k}} |z_{1,k}-\bar{z}_{1,k}+\rho\lambda_{z_{1,k}}|^2+|\tau_k-\bar{\tau}_k+\rho\lambda_{\bar{\tau}_k}|^2+|C_{1,k}-\bar{C}_{1,k}+\rho\lambda_{C_{1,k}}|^2 \tag{47a}$$

$$\text{s.t.}\quad \bar{\tau}_k+\frac{\left(\frac{\bar{C}_{1,k}}{L_k}-z_{1,k}\right)^2}{4}-\left(\frac{\left(\frac{\bar{C}_{1,k}}{L_k}+z_{1,k}\right)\left(\frac{\bar{C}^l_{1,k}}{L_k}+z^l_{1,k}\right)}{2}-\frac{\left(\frac{\bar{C}^l_{1,k}}{L_k}+z^l_{1,k}\right)^2}{4}\right)\leq 0. \tag{47b}$$

$$\bar{C}_{1,k}(\lambda_{4,k})=\frac{4L_k^2(C_{1,k}+\rho\lambda_{C_{1,k}})+\lambda_{4,k}(z_{1,k}L_k+\bar{C}^l_{1,k}+z^l_{1,k}L_k)}{4L_k^2+\lambda_{4,k}}, \tag{49}$$

$$\min_{\tilde{\tau}_k,\bar{R}_k,z_{2,k}} |z_{2,k}-\bar{z}_{2,k}+\rho\lambda_{z_{2,k}}|^2+|\tau_k-\tilde{\tau}_k+\rho\lambda_{\tilde{\tau}_k}|^2+|R_k-\bar{R}_k+\rho\lambda_{R_k}|^2 \tag{51a}$$

$$\text{s.t.}\quad \frac{L_k}{W}\tilde{\tau}_k+\frac{(R_k-z_{2,k})^2}{4}-\left(\frac{(R_k+z_{2,k})\left(R^l_k+z^l_{2,k}\right)}{2}-\frac{\left(R^l_k+z^l_{2,k}\right)^2}{4}\right)\leq 0. \tag{51b}$$

$$R_k(\lambda_{5,k})=\frac{(R^l_k+z^l_{2,k})\lambda^2_{5,k}+2(\bar{R}_k-\rho\lambda_{R_k}+\bar{z}_2-\rho\lambda_{z_{2,k}}+R^l_k+z^l_{2,k})\lambda_{5,k}+8(\bar{R}_k-\rho\lambda_{R_k})}{8+4\lambda_{5,k}}, \tag{53}$$

mentarity slackness condition is satisfied. Let us define

$$Q_k(\tilde{v}_{k,k'},\tilde{\mathbf{u}}_k,\bar{\phi}_k)\triangleq\sum_{k'\neq k}|\tilde{v}_{k,k'}|^2+\|\tilde{\mathbf{u}}_k\|^2\frac{\sigma^2}{P}-(\frac{\tilde{v}^{l*}_{k,k}\tilde{v}_{k,k}}{\bar{\phi}^l_k}+\frac{\tilde{v}^*_{k,k}\tilde{v}^l_{k,k}}{\bar{\phi}^l_k}-\frac{|\tilde{v}^l_{k,k}|^2\bar{\phi}_k}{\bar{\phi}^{l2}_k}). \tag{64}$$

When $Q_k(\tilde{v}_{k,k'}(0),\tilde{\mathbf{u}}_k(0),\bar{\phi}_k(0))\leq 0$, we have the optimal $\bar{\phi}_k=\bar{\phi}_k(0)$, $\tilde{\mathbf{u}}_k=\tilde{\mathbf{u}}_k(0)$, and $\tilde{v}_{k,k'}=\tilde{v}_{k,k'}(0)$, otherwise we must have $Q_k(\tilde{v}_{k,k'}(\lambda_{7,k}),\tilde{\mathbf{u}}_k(\lambda_{7,k}),\bar{\phi}_k(\lambda_{7,k}))=0$, which is equivalent to

$$\begin{aligned}\tilde{\mathbf{u}}^H_k(\lambda_{7,k})\mathbf{X}(\lambda_{7,k})\tilde{\mathbf{u}}_k(\lambda_{7,k})&-\tilde{\mathbf{u}}^H_k(\lambda_{7,k})\mathbf{m}(\lambda_{7,k})\\&-\mathbf{m}^H(\lambda_{7,k})\tilde{\mathbf{u}}_k(\lambda_{7,k})\\&+t_k(\lambda_{7,k})=0,\end{aligned} \tag{65}$$

where $\mathbf{X}(\lambda_{7,k})\triangleq\frac{\sigma^2}{P}\mathbf{I}+\frac{\mathbf{R}_k}{(1+\lambda_{7,k})^2}$, $\mathbf{m}(\lambda_{7,k})=\frac{\sum_{k'\neq k}\rho\lambda^*_{\tilde{v}_{k,k'}}\mathbf{H}_{k'}\mathbf{v}_{RF,k'}}{(1+\lambda_{7,k})^2}+\frac{\tilde{v}^{l*}_{k,k}}{\bar{\phi}^l_k}\mathbf{H}_k\mathbf{v}_{RF,k}$, and $t_k(\lambda_{7,k})=\frac{\sum_{k'\neq k}\rho^2\left|\lambda_{\tilde{v}_{k',k}}\right|^2}{(1+\lambda_{7,k})^2}+\frac{|\tilde{v}^l_{k,k}|^2}{\bar{\phi}^{l2}_k}\frac{2(\phi_k+\rho\lambda_{\phi_k})-\lambda_{7,k}\frac{|\bar{v}^l_{k,k}|^2}{\bar{\phi}^{l2}_k}}{2}+2\Re e\left\{\frac{\tilde{v}^{l*}_{k,k}}{\bar{\phi}^l_k}(\rho\lambda_{\tilde{v}_{k,k}}+\lambda_{7,k}\frac{\tilde{v}^l_{k,k}}{\bar{\phi}^l_k})\right\}$.

For the Hermitian matrix $\mathbf{R}_k$, we put forward the following decomposition:

$$\mathbf{R}_k=\tilde{\mathbf{V}}_k\mathbf{\Lambda}_k\tilde{\mathbf{V}}^H_k, \tag{66}$$

where $\tilde{\mathbf{V}}_k$ denotes a unitary matrix which contains the eigenvectors of $\mathbf{R}_k$, and $\mathbf{\Lambda}_k$ is a diagonal matrix consisting of the eigenvectors of $\mathbf{R}_k$. Then we have

$$\mathbf{A}^{-1}(\lambda_{7,k})=\tilde{\mathbf{V}}_k\left[\left(1+\lambda_{7,k}\frac{\sigma^2}{P}\right)\mathbf{I}+\mathbf{\Lambda}_k\frac{\lambda_{7,k}}{1+\lambda_{7,k}}\right]^{-1}\tilde{\mathbf{V}}_k. \tag{67}$$

Similarly, $\mathbf{X}(\lambda_{7,k})$ can be rewritten as

$$\mathbf{X}(\lambda_{7,k})=\tilde{\mathbf{V}}_k\left(\frac{\sigma^2}{P}\mathbf{I}+\frac{\mathbf{\Lambda}_k}{(1+\lambda_{7,k})^2}\right)\tilde{\mathbf{V}}^H_k. \tag{68}$$

Let $\mathbf{Y}_k(\lambda_{7,k})\triangleq(1+\lambda_{7,k}\frac{\sigma^2}{P})\mathbf{I}+\mathbf{\Lambda}_k\frac{\lambda_{7,k}}{1+\lambda_{7,k}}$ and $\mathbf{Z}_k(\lambda_{7,k})\triangleq\frac{\sigma^2}{P}\mathbf{I}+\frac{\mathbf{\Lambda}_k}{(1+\lambda_{7,k})^2}$. Then (65) can be rewritten as

$$\begin{aligned}&\mathrm{Tr}\left\{\mathbf{Y}^{-1}_k\mathbf{Z}_k\mathbf{Y}^{-1}_k\tilde{\mathbf{V}}^H_k\mathbf{b}\mathbf{b}^H\tilde{\mathbf{V}}_k\right\}\\&+\mathrm{Tr}\left\{\mathbf{Y}^{-1}_k\left[\tilde{\mathbf{V}}^H_k(\mathbf{Z}_k\mathbf{b}^H+\mathbf{b}\mathbf{Z}^H_k)\tilde{\mathbf{V}}_k\right]\right\}+t_k=0.\end{aligned} \tag{69}$$

Finally, (69) can be equivalently expressed as

$$\begin{aligned}&\sum_{i=1}^{N_{\mathrm{BS}}}[\mathbf{Y}^{-1}_k\mathbf{Z}_k\mathbf{Y}^{-1}_k]_{i,i}[\tilde{\mathbf{V}}^H_k\mathbf{b}\mathbf{b}^H\tilde{\mathbf{V}}_k]_{i,i}\\&+\sum_{i=1}^{N_{\mathrm{BS}}}[\mathbf{Y}^{-1}_k]_{i,i}[\tilde{\mathbf{V}}^H_k(\mathbf{Z}_k\mathbf{b}^H+\mathbf{b}\mathbf{Z}^H_k)\tilde{\mathbf{V}}_k]_{i,i}+t_k=0.\end{aligned} \tag{70}$$

Since $\mathbf{Y}_k$ and $\mathbf{Z}_k$ are both diagonal matrices, (70) can be easily solved using one dimensional search. Finally, by substituting the optimal $\lambda_{7,k}$, we obtain the solution for $\{\tilde{v}_{k,k'},\tilde{\mathbf{u}}_k,\bar{\phi}_k\}$.

C. Variables Updating in Step 3

In **Step 3**, we update t and $\mathbf{U}_{RF}$ in parallel by fixing the other block of variables. To this end, the subproblem with

$$\lambda_{6,k}^{\star} = \max\left\{0, \frac{2(L_k^2 - L_k^2(\tau_k + \rho\lambda_{\hat{\tau}_k}) - C_{2,k}L_k(t + \rho\lambda_{\bar{t}_k}))}{L_k^2 + C_{2,k}^2}\right\}. \tag{58}$$

$$\min_{\tilde{v}_{k,k'},\tilde{\mathbf{u}}_k,\bar{\phi}_k} \sum_{k'=1}^{K} |\tilde{v}_{k,k'} - \tilde{\mathbf{u}}_k^H \mathbf{H}_{k'} \mathbf{v}_{RF,k'} + \rho\lambda_{\tilde{v}_{k,k'}}|^2 + |\phi_k - \bar{\phi}_k + \rho\lambda_{\phi_k}|^2 + \|\tilde{\mathbf{u}}_k - \mathbf{U}_{RF}\mathbf{u}_{BB,k} + \rho\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k}\|^2 \tag{59a}$$

$$\text{s.t.} \quad \sum_{k'\neq k} |\tilde{v}_{k,k'}|^2 + \|\tilde{\mathbf{u}}_k\|^2 \frac{\sigma^2}{P} - \left(\frac{\tilde{v}_{k,k}^{l*}\tilde{v}_{k,k}}{\bar{\phi}_k^l} + \frac{\tilde{v}_{k,k}^{*}\tilde{v}_{k,k}^{l}}{\bar{\phi}_k^l} - \frac{|\tilde{v}_{k,k}^{l}|^2\bar{\phi}_k}{\bar{\phi}_k^{l2}}\right) \leq 0. \tag{59b}$$

respect to t can be solved in closed-form solutions by taking advantage of the first-order optimality condition while $\mathbf{U}_{RF}$ can be solved in the one-iteration BCD fashion.

1) **Subproblem for** t: The subproblem with respect to t is an unconstrained quadratic optimization problem, which is shown as

$$\min_{t} \; t + \frac{1}{2\rho}\sum_{k=1}^{K}\left(|t - t_k + \rho\lambda_{t_k}|^2 + |t - \bar{t}_k + \rho\lambda_{\bar{t}_k}|^2\right). \tag{71}$$

By examining the first order optimality condition, the optimal t is derived as

$$t = \frac{-\rho + \sum_{k=1}^{K}(t_k - \rho\lambda_{t_k} + \bar{t}_k - \rho\lambda_{\bar{t}_k})}{2K}. \tag{72}$$

2) **Subproblem for** $\mathbf{U}_{RF}$: The subproblem with respect to $\mathbf{U}_{RF}$ is a quadratic optimization problem with unit modulus constraints, which is given by

$$\min_{\mathbf{U}_{RF}} \sum_{k=1}^{K} \|\tilde{\mathbf{u}}_k - \mathbf{U}_{RF}\mathbf{u}_{BB,k} + \rho\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k}\|^2 \tag{73a}$$

$$\text{s.t.} \quad |\mathbf{U}_{RF}(i,j)| = 1, \quad \forall i,j. \tag{73b}$$

Similar to the processing method in dealing with $\{\mathbf{v}_{RF,k}\}$, we rewrite the subproblem as follows

$$\min_{\mathbf{U}_{RF}} \; \mathrm{Tr}\left(\mathbf{U}_{RF}^H\tilde{\mathbf{C}}\mathbf{U}_{RF}\tilde{\mathbf{P}}\right) - 2\Re e\left\{\mathrm{Tr}\left(\mathbf{U}_{RF}^H\tilde{\mathbf{Q}}\right)\right\} \tag{74a}$$

$$\text{s.t.} \quad |\mathbf{U}_{RF}(i,j)| = 1, \quad \forall i,j. \tag{74b}$$

where $\tilde{\mathbf{C}} = \mathbf{I}$, $\tilde{\mathbf{P}} = \sum_{k=1}^{K}\mathbf{u}_{BB,k}\mathbf{u}_{BB,k}^H$, and $\tilde{\mathbf{Q}} = \sum_{k=1}^{K}(\tilde{\mathbf{u}}_k + \rho\boldsymbol{\lambda}_{\tilde{\mathbf{u}}_k})\mathbf{u}_{BB,k}^H$. Again, it can be solved by using the same method in [42, Appendix B].

Appendix C
Proposed CM-based heuristic algorithm

We propose a heuristic algorithm for comparison. In this algorithm, the analog beamformers are designed based on the channel matching (CM) method [22], the digital beamformer $\mathbf{U}_{BB}$ is designed according to the conventional minimum mean squared error (MMSE) design criterion, and the resource allocation is in a heuristic way. To be specified, we calculate the rank one truncated SVD of $\mathbf{H}_k$ as

$$\mathbf{H}_k \approx \mathbf{U}_{s,k}\boldsymbol{\Sigma}_{s,k}\mathbf{v}_{s,k}^H, \tag{75}$$

where $\mathbf{U}_{s,k} \in \mathbb{C}^{N_{\mathrm{BS}}\times N_{\mathrm{BS},k}}$, $\mathbf{v}_{s,k} \in \mathbb{C}^{N_{\mathrm{user}}\times 1}$, and $\boldsymbol{\Sigma}_{s,k} \in \mathbb{C}^{N_{\mathrm{BS},k}\times 1}$ with $\sum_{k=1}^{K} N_{\mathrm{BS},k} = N_{\mathrm{s}}$. Then the analog beamformers $\mathbf{U}_{RF}$ and $\mathbf{v}_{RF,k}$ can be expressd as

$$\mathbf{U}_{RF} = [e^{j\angle(\mathbf{U}_{s,1})}, ..., e^{j\angle(\mathbf{U}_{s,K})}], \tag{76a}$$

$$\mathbf{v}_{RF,k} = e^{j\angle(\mathbf{v}_{s,k})}, \tag{76b}$$

where the operator $\angle(\mathbf{A})$ computes the angle of $\mathbf{A}$ element-wise. Based on the MMSE technique, the digital beamformer $\mathbf{U}_{BB}$ can be obtained as

$$\mathbf{U}_{BB} = \mathbf{T}\left(\mathbf{T}^H\mathbf{T} + K\frac{\sigma^2}{P}\mathbf{I}\right)^{-1}, \tag{77}$$

where $\mathbf{T} = \sum_{k=1}^{K}\mathbf{U}_{RF}^H\mathbf{H}_k\mathbf{v}_{RF,k}$.

The resource is allocated in a heuristic way. First, the computational resources allocated to user k is proportional to the amount of its task, i.e., $\frac{C_{1,1}}{L_1} = , ..., = \frac{C_{1,K}}{L_K}$, with $\sum_{k=1}^{K} C_{1,k} = C_{\max}$. Then the offloading ratio of user k is obtained by $\tau_k = \frac{C_{1,k}}{C_{1,k}+C_{2,k}}$.

References

[1] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing: A key technology towards 5G," *ETSI White Paper* no. 11, Sep. 2015.

[2] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: new paradigms, scenarios, and challenges," *IEEE Commun. Mag.,* vol. 55, no. 4, pp. 54–61, Apr. 2017.

[3] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Survey Tuts.,* vol. 19, no. 4, pp. 2322–2358, Aug. 2017.

[4] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.,* vol. 64, no. 10, pp. 4268–4282, Oct. 2016

[5] C. You, K. Huang and H. Chae, "Energy efficient mobile cloud computing powered by wireless energy transfer," *IEEE J. Sel. Areas Commun.,* vol. 34, no. 5, pp. 1757–1771, May. 2016.

[6] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT),* Barcelona, Spain, Jul. 2016, pp. 1451–1455.

[7] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.,* vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[8] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.,* vol. 16, no. 3, pp. 1397–1411, Mar. 2017.

[9] X. Lyu, H. Tian, P. Zhang, and C. Sengul, "Multi-user joint task offloading and resources optimization in proximate clouds," *IEEE Trans. Veh. Techn.,* vol. 66, no. 4, pp. 3435–3447, Apr. 2017.

[10] J. Ren, G. Yu, Y. Cai, and Y. He, "Latency optimization for resource allocation in mobile-edge computation offloading," *IEEE Trans. Wireless Commun.,* vol. 17, no. 8, pp. 5506–5519, Aug. 2018.

[11] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Power-delay tradeoff in multi-user mobile-edge computing systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM),* Washington, DC, USA, Dec. 2016, pp. 1–6.

[12] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *EEE/ACM Trans. Networking,* vol. 24, no. 5, pp. 2795–2808., Oct. 2016.

[13] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.,* vol. 49, no. 6, pp. 101–107, Jun. 2011.

[14] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimo, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access,* vol. 1, pp. 335–349, May 2013.

[15] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capability evaluation," *IEEE J. Sel. Areas Commun.,* vol. 32, no. 6, pp. 1164–1179, June. 2014.

[16] V. Frascolla, F. Miatton, G. K. Tran, K. Takinami, A. D. Domenico, E. C. Strinati, K. Koslowski, T. Haustein, K. Sakaguchi, S. Barbarossa, and S. Barberis, "5G-MiEdge: design, standardization and deployment of 5G Phase II technologies," in *IEEE Conference on Standards for Communications and Networking (CSCN),* Helsinki, Finland, Sep. 2017, pp. 1–6.

[17] F. Haider, X. Gao, X. H. You, Y. Yang, D. Yuan, H. M. Aggoune, and H. Haas, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.,* vol. 52, no. 2, pp. 122–130, Feb. 2014.

[18] E. Torkildson, U. Madhow, and M. Rodwell, "Indoor Millimeter Wave MIMO: Feasibility and Performance," *IEEE Trans. wireless Commun.*, vol. 10, no. 12, pp. 4150–4160, Dec. 2011.

[19] V. Venkateswaran and A. J. van der Veen, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Trans. Signal Process.,* vol. 58, no. 8, pp. 4131–4143, Aug. 2010.

[20] Y. Cai, Y. Xu, Q. Shi, B. Champagne and L. Hanzo, "Robust joint hybrid transceiver design for millimeter wave full-duplex MIMO relay systems," *IEEE Trans. Wireless Commun.,* vol. 18, no. 2, pp. 1199–1215, Feb. 2019.

[21] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.,* vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[22] J. Zhang, M. Haardt, I. Soloveychik, and A. Wiesel, "A channel matching based hybrid analog-digital strategy for massive multi-user MIMO downlink systems," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop,* Jul. 2016, pp. 1–5.

[23] X. Zhai, Y. Cai, Q. Shi, M. Zhao, G. Y. Li, and B. Champagne, 'Joint transceiver design with antenna selection for large-scale MU-MIMO millimeter-wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2085–2096, Sep. 2017.

[24] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing,* vol. 10, no. 3, pp. 501–513, Apr. 2016.

[25] 3GPP TR 22.891, "Feasibility Study on New Services and Markets Technology Enablers," Ver. 14.2.0, Sep. 2016.

[26] S. Barbarossa, E. Ceci, M. Merluzzi and E. Calvanese, "Enabling effective mobile edge computing using millimeterwave links," in *IEEE International Commun. Conf. (ICC),* Paris, France, May, 2017, pp. 1–6.

[27] Q. Shi, M. Hong, X. Fu, T.-H. Chang, "Penalty dual decomposition method for nonsmooth nonconvex optimization," [Online]. Available: https://arxiv.org/abs/1712.04767

[28] Q. Shi and M. Hong, "Penalty dual decomposition method with application in signal processing," in *Proc. Int. Conf. on Acoust. Speech Signal Process. (ICASSP),* New Orleans, USA, Mar., 2017, pp. 4059–4063.

[29] 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.

[30] M. M. Zhao, Q. Shi, Y. Cai, M. J. Zhao and Y. Li, "Distributed Penalty Dual Decomposition Algorithm for Optimal Power Flow in Radial Networks," *IEEE Trans. Power Syst.,* DOI: 10.1109/TPWRS.2019.2952433, Nov. 2019.

[31] D. Bertsekas, *Nonlinear Programming,* 2nd ed. Belmont, MA: Athena Scientific, 1999.

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

[33] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, Apr. 2003.

[34] G. R. Lanckriet and B. K. Sriperumbudur, "On the convergence of the concave-convex procedure," in *Advances in Neural Info. Processing Systems*, pp. 1759–1767, 2009.

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

[36] M. Hong, T.-H. Chang, X. Wang, M. Razaviyayn, S. Ma, and Z.-Q. Luo, "A block successive upper bound minimization method of multipliers for linearly constrained convex optimization," [Online]. Available: https://arxiv.org/abs/1401.7079

[37] G. R. MacCartney, Jr., T. S. Rappaport, S. Sun, and S. Deng, "Indoor office wideband millimeter-wave propagation measurements and channel models at 28 and 73 GHz for ultra-dense 5G wireless networks," *IEEE Access,* vol. 3, pp. 2388-2424, Oct. 2015.

[38] P. Smulders and L. Correia, "Chracterization of propagation in 60 GHz radio channels," *Electron. Commun. Eng. J.,* vol. 9, no. 2, pp. 73–80, Apr. 1997.

[39] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," *in Proc. USENIX Conf. Hot Topics Cloud Comput. (HotCloud)*, Boston, MA, Jun. 2010, pp. 1–7.

[40] S. Melendez and M. P. McGarry, "Computation offloading decisions for reducing completion time," *in IEEE Annual Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, Jan. 2017, pp. 160–164.

[41] C. Zhao, Y. Cai, and M. Zhao, "Joint hybrid beamforming and offloading for mmWave mobile edge computing systems," in *Proc. of the IEEE WCNC*, Marrakech, Morocco, Apr. 2019, pp. 1–5.

[42] Q. Shi, and M. Hong, "Spectral efficiency optimization for mmWave multiuser MIMO systems," *IEEE J. Sel. Topics Signal Process.,* vol. 12, no. 3, pp. 455–468, Jun. 2018.