Edge and Central Cloud Computing: A Perfect Pairing for High Energy Efficiency and Low-Latency

Xiaoyan Hu, *Student Member, IEEE*, Lifeng Wang, *Member, IEEE*, Kai-Kit Wong, *Fellow, IEEE*, Meixia Tao, *Fellow, IEEE*, Yangyang Zhang, and Zhongbin Zheng

***Abstract*— In this paper, we study the coexistence and synergy between edge and central cloud computing in a heterogeneous cellular network (HetNet), which contains a multi-antenna macro base station (MBS), multiple multi-antenna small base stations (SBSs) and multiple single-antenna user equipment (UEs). The SBSs are empowered by edge clouds offering limited computing services for UEs, whereas the MBS provides high-performance central cloud computing services to UEs via a restricted multiple-input multiple-output (MIMO) backhaul to their associated SBSs. With processing latency constraints at the central and the edge networks, we aim to minimize the system energy consumption used for task offloading and computation. The problem is formulated by jointly optimizing the cloud selection, the UEs' transmit powers, the SBSs' receive beamformers, and the SBSs' transmit covariance matrices, which is a mixed-integer and non-convex optimization problem. Based on the methods such as decomposition approach and successive pseudoconvex approach, a tractable solution is proposed via an iterative algorithm. The simulation results show that our proposed solution can achieve great performance gain over conventional schemes using edge or central cloud alone. Also, with large-scale antennas at the MBS, the massive MIMO backhaul can significantly reduce the complexity of the proposed algorithm and obtain even better performance.**

***Index Terms*— Edge computing, central cloud computing, HetNets, backhaul, massive MIMO.**

Manuscript received November 8, 2018; revised May 17, 2019 and August 27, 2019; accepted October 22, 2019. Date of publication November 7, 2019; date of current version February 11, 2020. This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K015893/1. This article is to be presented in part at the IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, December 2019. The associate editor coordinating the review of this article and approving it for publication was S. Mukherjee. *(Corresponding author: Xiaoyan Hu.)*

X. Hu and K.-K. Wong are with the Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K. (e-mail: xiaoyan.hu.16@ucl.ac.uk; kai-kit.wong@ucl.ac.uk).

L. Wang is with the Department of Electrical Engineering, Fudan University, Shanghai 200433, China (e-mail: lifengwang@fudan.edu.cn).

M. Tao is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: mxtao@sjtu.edu.cn).

Y. Zhang is with the Kuang-Chi Institute of Advanced Technology, Shenzhen 518057, China (e-mail: yangyang.zhang@kuang-chi.org).

Z. Zheng is with the East China Institute of Telecommunications, China Academy of Information and Communications Technology, Shanghai 200001, China (e-mail: ben@ecit.org.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.2019.2950632

I. Introduction

A. Motivations

TO DEAL with the computation-intensive tasks generated by a large variety of applications, the concept of central cloud computing (CCC) first emerges, which offloads these tasks to remote powerful computing centers, also known as central clouds. Mobile cloud computing (MCC) integrates CCC into mobile environment to facilitate mobile users to take full advantages of cloud resources [1], [2]. Although CCC/MCC can provide high-performance computing services for UEs, it has one inherent disadvantage, i.e., the central clouds are usually located far away from UEs. Hence, accessing the CCC/MCC services induces excessive transmission latency, which aggravates the backhaul congestion. Besides, it is easy to encounter the performance bottleneck considering the finite backhaul capacity and exponentially growing mobile data, which has led to the emergence of edge (cloud) computing.

Edge (cloud) computing has recently been regarded as one of the key enablers to shape the future intelligent wireless networks. The rationale behind edge computing is that cloud computing can be carried out at the edge of wireless networks which is close to UEs, so as to facilitate computation offloading of UEs and prolong their battery lifetime [3], [4]. The standardization bodies and industry associations such as ETSI and 5GAA have identified various edge computing use cases for 5G cellular networks, such as vehicle-to-everything (V2X) and massive machine-type communications (mMTC), etc., [5], [6]. For practical deployment, several edge computing architectures have already been proposed, such as mobile edge computing (MEC) [7], fog computing [8], [9], and also cloudlets [10].

Edge computing is of great benefit to resource-limited UEs, e.g., the Internet-of-Things (IoT) devices, which avoids the frequent delivery of massive computing tasks to the core networks with central cloud for computing, and thus reduces the transmission latency and backhaul congestion [11], [12]. However, the computing capabilities at the edge servers/clouds are also limited in general due to the cost and their constrained sizes. For UEs with highly computation-intensive tasks, the edge computing servers/clouds may be incapable of providing them

with satisfactory computing services. Under this situation, CCC/MCC has been shown to be an effective solution. Hence, in order to improve the quality of service (QoS) for dealing with a wide range of UEs' computation tasks, applying the architecture with the coexistence and cooperation between the edge and central clouds could be a promising option.

B. Related Works

Extensive works on CCC/MCC have been conducted to explore the potential of central cloud. Several system architectures with various code offloading frameworks have been studied, e.g., MAUI [13] and ThinkAir [14]. In [15], dynamic resource allocation using virtualization technology was studied to achieve overload avoidance and green computing by minimizing the number of physical machines. Also, a computation offloading algorithm was proposed in [16] to deal with multiple services in workflow by leveraging MCC.

Recently, considerable attention has been paid to the design and analysis of edge computing in cellular networks, e.g., [17]–[21]. The tradeoff between energy consumption and latency in information transmission and computation was analyzed in [17], where an UE offloaded its application tasks to a small BS (SBS) for processing. In [18], a multi-user computation offloading problem was considered in a single-cell scenario and game-theoretical solutions were proposed in order to maximize the cell load and minimize the cost in terms of computational time and energy simultaneously. Later in [19], time and frequency allocation problems for improving energy efficiency were studied by considering multi-user computation offloading in a single cell equipped with limited cloud capacity, where an offloading priority function was derived to accommodate users' priorities. The work of [20] examined a single-cloudlet scenario where multiple UEs were served with equal-time sharing, and a successive convex optimization approach was developed to minimize the network energy consumption under a computing latency constraint. Recent works related to edge computing also focus on multi-service scenarios. For example, [21] considered a single MEC server with storage capability and attempted to maximize the revenue of providing both the computing and caching services. Besides, the effects of implementing the technology of edge computing in energy harvesting networks have been verified in [22]–[26], where the UEs' battery lifetime could be further prolonged.

The complementary benefits between edge and central clouds have driven research towards the coexistence and cooperation between edge and central clouds [27]. One such example was [28] where a delay-aware scheduling between local and Internet clouds was studied, and a priority-based cooperation policy was given to maximize the total successful offloading probability. The placement and provisioning of virtualized network functions were explored in [29], in which a QoS-aware optimization strategy was proposed over an edge-central carrier cloud infrastructure. Also, the work in [30] considered that an edge server and a central cloud coexist to complete the UEs' computations cooperatively, where a wired connection was assumed between the edge and the central clouds.

C. Our Contributions

Most of the existing computing works focused on either the edge or central cloud computing independently, and the edge computing works mainly concentrated on small-scale networks such as the single MEC server or cloudlet case [12], [18]–[21], [23]–[26]. Even though edge computing has been regarded as a promising trend to deal with the ever growing mobile computing data, it cannot entirely replace the present central cloud computing, due to the fact that edge computing is set to push limited processing and storage capabilities close to UEs but may be incapable of dealing with big data processing. The latest white paper from ETSI has further illustrated that central cloud computing and edge computing are highly complementary and significant benefits can be attained when utilizing them both [5]. However, the existing works [28]–[30] considering the coexistence of edge and central cloud computing either focus on delay-aware priority scheduling, virtualized resource allocation, or offloading with wired backhaul. The issues related to offloading decision and resource allocation of hybrid edge/central cloud computing networks with wireless backhaul have not been thoroughly studied, especially from the viewpoint of communication [4]. Therefore, this paper studies the deployment of heterogeneous edge and central clouds to leverage the easy access of edge clouds and the abundant computing resources at the central cloud, mainly from the viewpoint of communication by considering cloud selection and resource allocation. To our best knowledge, this is the first work addressing the integrated edge and central cloud computing in heterogeneous cellular networks (HetNets) while considering the physical properties of wireless backhaul.

Our main contributions are summarized as follows:

- **Hybrid Edge/Central Cloud Computing Architecture:** We consider a hybrid edge and central cloud computing architecture in a two-tier HetNet, including one macro cell with a macro BS (MBS) and multiple small cells each with an SBS. The edge clouds with limited computing capabilities are co-located at or linked to the SBSs by error-free optical fibers while the central cloud with ultra-high computing capability is connected with the MBS through optical fibers as well. The UEs can offload their computation tasks directly to the SBSs to access the edge cloud computing services (edge computing mode) or further offload to the MBS through the restricted multiple-input multiple-output (MIMO)/massive MIMO backhaul to utilize the central cloud computing services (central computing mode). Cooperation of edge and central clouds will improve the QoS and ensure the scalability and load balancing between the edge and central clouds.
- **Problem Formulation with Joint Optimization on the Cloud Selection, Access Transmit Powers, Receive Beamforming Vectors and Backhaul Transmit Covariance Matrices:** Our aim is to minimize the network's energy consumption for task offloading and computation under both the central and edge processing latency constraints through jointly optimizing the cloud selection, the UEs' transmit powers, the SBSs' receive beamforming vectors, and the SBSs' transmit covariance matrices.

The central processing latency constraints require the backhaul transmission latencies being lower than the corresponding computing latencies at the edge clouds; otherwise, the central cloud will not be selected. The edge processing latency constraints require the corresponding latencies not exceeding a targeted threshold to guarantee the quality of services provided by the edge clouds. A mixed-integer and non-convex optimization problem is formulated accordingly, which is NP-hard in general. For the case of massive MIMO backhaul, we consider two low-complexity linear processing methods, namely maximal-ratio combining (MRC) and zero-forcing (ZF), and the corresponding optimization problems can be much simplified.

- **Algorithm Design:** An iterative algorithm is developed to solve the combinatorial mixed-integer and non-convex optimization problem corresponding to the case with traditional MIMO backhaul. In particular, we show that in each iteration, the UEs' transmit powers and the SBSs' receive beamforming vectors can be optimized in closed-form, and the SBSs' transmit covariance matrix solution is obtained by leveraging a successive pseudoconvex optimization approach. In addition, the massive MIMO backhaul solutions can be easily obtained thanks to the unique features of massive MIMO transmission, which significantly reduce the complexity of the algorithm. The practicality of the proposed algorithm lies in that it can properly address the issues of cloud selection and resource allocation for a HetNet architecture with hybrid edge/central cloud computing resources while considering the physical properties of wireless backhaul.
- **Design Insights:** Simulation results are presented to demonstrate the efficiency of the proposed algorithm and shed light on the effects of key parameters such as the offloaded task size, edge processing latency threshold, and edge clouds' CPU frequency. It is confirmed that the solution of the integrated edge and central cloud computing scheme proposed in this work can achieve better performance than the schemes with edge (cloud) computing alone or central cloud computing alone, and outperforms all the other benchmark solutions. In addition, low-complexity massive MIMO solution with ZF receiver could always outperform the solution with traditional MIMO backhaul, while the solution with MRC receiver could achieve similar or better performance than the traditional MIMO one in certain scenarios.

The rest of this paper is organized as follows. In Section II, the considered system model is described and the corresponding optimization problem is formulated. The proposed algorithm under traditional MIMO backhaul is presented in Section III, and massive MIMO backhaul solutions are given in Section IV. Section V provides the simulation results. Finally, we have some concluding remarks in Section VI.

Notations—In this paper, the upper and lower case bold symbols denote matrices and vectors, respectively. The notations $(\cdot)^H$ and $(\cdot)^\dagger$ are conjugate transpose and conjugate operators for vectors or matrices, respectively. $[x]^+$ is used as $\max\{x, 0\}$. In addition, $\det(\mathbf{A})$ denotes the determinant of $\mathbf{A}$, and $\mathrm{tr}\{\mathbf{A}\}$ is the trace of $\mathbf{A}$. Also, $\mathrm{eig}\{\mathbf{A}\}$ denotes the set of all the eigenvalues for $\mathbf{A}$, and $\mathrm{eigvec}\{\cdot\}$ gives the eigenvector for a given eigenvalue of $\mathbf{A}$. $\langle \mathbf{A}_1, \mathbf{A}_2 \rangle \triangleq \Re\{\mathrm{tr}(\mathbf{A}_1^H \mathbf{A}_2)\}$, where $\Re\{\cdot\}$ is the real-value operator, and $\nabla_{\mathbf{A}} f(\mathbf{A})$ denotes the Jacobian matrix of $f(\mathbf{A})$ with respect to (w.r.t.) $\mathbf{A}$.

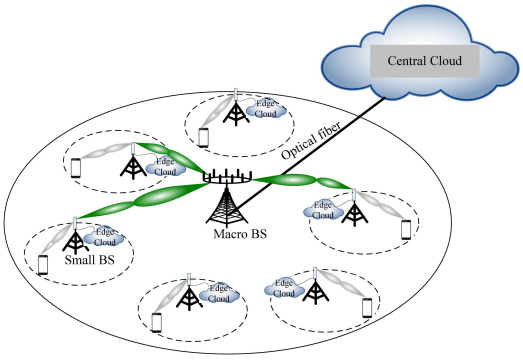


Fig. 1. An illustration of two-tier HetNets equipped with edge clouds associated with the SBSs and central cloud connected by the MBS via optical fiber, where the MBS provides central cloud computing services for UEs through restricted MIMO/massive MIMO backhaul for addressing more complicated computing tasks which cannot be efficiently handled by the SBSs' edge clouds due to the limited computing capabilities.

II. System Model and Problem Formulation

As shown in Fig. 1, we consider a two-tier HetNet, where an M-antenna MBS provides wireless MIMO backhaul and is fiber-optic connected to the central cloud with super computing capability, and N SBSs with edge clouds can provide limited computing capabilities.[1] In each small cell, an SBS equipped with L antennas serves a single-antenna UE.[2] Note that existing user association schemes [32] can be adopted to determine which user is connected to an SBS. Let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of the SBSs as well as the UEs. Each UE $n \in \mathcal{N}$ has an atomic highly integrated computation-intensive task, characterized by a positive tuple $[I_n, O_n, K_n]$, which cannot be partitioned for parallel execution. Here I_n is the size (in bits) of the computation task-input data (e.g., the program codes and the input parameters) which cannot be divided and has to be offloaded as a whole for computing, O_n denotes the size of the task-output data corresponding to the results of the I_n input data, and K_n is the amount of required computing resources for computing 1-bit of the input data (i.e., the number of CPU cycles required).[3]

[1]The central cloud can be regarded as the computing part of the cloud radio access network (Cloud RAN) [31]. Each edge cloud can be an independent edge computing server co-located at the corresponding SBS or a certain part of computing capability allocated to the SBS from a nearby fiber-optic connected edge computing center [4].

[2]The extended case of serving multiple UEs in each small cell can be effectively dealt with by using the existing orthogonal multiple access techniques for radio resource allocation. In addition, the extended case of our work can be viewed as leveraging equal computing resource sharing at a SBS for multiple active UEs in a small cell, or dedicated computing resource policy for different types of computing services, i.e., each service will be granted one dedicated computing resource.

[3]The parameters in the task tuple of $[I_n, O_n, K_n]$ can be obtained through task profilers by applying the methods (e.g., call graph analysis) [4], [13], [33]–[35]. It is assumed that the size of computing outputs, i.e., O_n (usually a few command bits) is much smaller than I_n (usually measured by Mbit) in practice, and thus the downlink overhead such as time and energy consumption for delivering the output data back to the UEs is negligible and can be ignored.

Let B^{a} and B^{b} denote the bandwidths allocated to UEs' access links to their serving SBSs and SBSs' backhaul links to the MBS, respectively. A coordination and monitoring protocol between SBSs and MBS, as the one used in [36], [37], is needed.

Assuming that the UEs are endowed with very limited computing resources, they tend to choose computation offloading to complete their computation tasks remotely, so as to save their own energy and resources. Since the computation tasks offloaded by the UEs could be executed either at the edge clouds or central cloud, the cloud selection needs to be appropriately determined before evaluating the computation latency and energy consumption. Let the binary indicator c_n denote the computing decision, where $c_n = 1$ indicates edge computing, and $c_n = 0$ indicates central cloud computing being selected for each UE $n \in \mathcal{N}$. In the sequel, we will study the latency and energy consumption of the network, and then formulate the optimization problem for minimizing the network's total energy consumption for task offloading and computation under the central and edge processing latency constraints.

A. Transmission and Computing Latency

1) Access Transmission Latency: The uplink transmission rate of UE n for offloading the I_n-bit computation tasks to its serving SBS n is expressed as

$$R_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) = B^{\mathrm{a}} \log_2 \left(1 + \gamma_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n)\right), \quad n \in \mathcal{N} \tag{1}$$

with the signal-to-interference-plus-noise ratio (SINR)

$$\gamma_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) = \frac{p_n^{\mathrm{u}} |\mathbf{w}_n^H \mathbf{h}_{n,n}^{\mathrm{a}}|^2}{\sum_{i=1, i\neq n}^{N} p_i^{\mathrm{u}} |\mathbf{w}_n^H \mathbf{h}_{i,n}^{\mathrm{a}}|^2 + |\mathbf{w}_n^H \mathbf{n}_n|^2}, \tag{2}$$

where $\mathbf{w}_n$ is the receive beamforming vector of the n-th SBS, $\mathbf{h}_{i,n}^{\mathrm{a}} \in \mathbb{C}^{L\times 1}$ is the access channel vector between UE i and SBS n, $\mathbf{n}_n$ is a vector of the additive white Gaussian noise with zero mean and variance σ_n^2, and $\mathbf{p}^{\mathrm{u}} \triangleq [p_1^{\mathrm{u}}, \ldots, p_N^{\mathrm{u}}]^T \in \mathbb{R}^{N\times 1}$ denotes the transmit power vector of the UEs. Therefore, the uplink access transmission latency for offloading UE n's task can be calculated as

$$T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) = \frac{I_n}{R_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n)}, \quad n \in \mathcal{N}. \tag{3}$$

2) Edge Computing Latency ($c_n = 1$): Let f_n denote the CPU clock frequency of the n-th edge cloud associated with SBS n, and thus the corresponding edge computation latency for dealing with the I_n-bits input data can be described as

$$T_n^{\mathrm{edge}} = \frac{I_n K_n}{f_n}, \quad n \in \mathcal{N}, \tag{4}$$

which indicates that the value of edge computing latency heavily depends on the offloaded task size, the required unit computing resources and edge cloud's CPU clock frequency.

3) Central Cloud Processing Latency/Backhaul Transmission Latency ($c_n = 0$): The central cloud processing latency mainly comes from backhaul transmission and task execution at the central cloud. Due to the central cloud's super computing capability, its computing time is much lower than edge computing, thus we assume that the time for central cloud computing is negligible. Hence, the central cloud processing latency, i.e., the backhaul transmission latency, for the n-th UE can be calculated as[4]

$$T_n^{\mathrm{central}}(\mathbf{Q}) = \frac{I_n}{R_n^{\mathrm{b}}(\mathbf{Q})}, \quad n \in \mathcal{N}, \tag{5}$$

where $R_n^{\mathrm{b}}(\mathbf{Q})$ is the backhaul transmission rate given by

$$R_n^{\mathrm{b}}(\mathbf{Q}) = B^{\mathrm{b}} \log_2 \det \left(\mathbf{I} + \Psi(\mathbf{Q}_{-n})^{-1} \mathbf{H}_n^{\mathrm{b}} \mathbf{Q}_n \left(\mathbf{H}_n^{\mathrm{b}}\right)^H\right), \tag{6}$$

with the noise-plus-interference covariance matrix $\Psi(\mathbf{Q}_{-n}) = \sigma^2 \mathbf{I} + \sum_{i=1, i\neq n}^{N} \mathbf{H}_i^{\mathrm{b}} \mathbf{Q}_i \left(\mathbf{H}_i^{\mathrm{b}}\right)^H$. In (6), $\mathbf{Q}_n$ is the transmit covariance matrix of SBS n, $\mathbf{Q} = \{\mathbf{Q}_n\}_{n=1}^N$ and $\mathbf{Q}_{-n} = \{\mathbf{Q}_i\}_{i=1, i\neq n}^N$ are respectively the compact transmit covariance matrices and the compact transmit covariance matrices except $\mathbf{Q}_n$, and $\mathbf{H}_n^{\mathrm{b}} \in \mathbb{C}^{M\times L}$ is the backhaul channel matrix from SBS n to the MBS. Note that if the task of UE $n \in \mathcal{N}$ is executed by the edge cloud of SBS n, i.e. $c_n = 1$, the transmit covariance matrix at SBS n shall be $\mathbf{Q}_n = \mathbf{0}$.

B. Energy Consumption

Network energy consumption mainly results from task offloading and task execution/computation. Based on Section II-A, the amount of energy consumption of UE $n \in \mathcal{N}$ for offloading its computation task to its serving SBS n can be calculated as

$$E_n^{\mathrm{a}} = p_n^{\mathrm{u}} T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n), \quad n \in \mathcal{N}. \tag{7}$$

If the UE n's task is executed by the edge cloud associated with the SBS n, the computation energy consumption at the corresponding edge server is given by

$$E_n^{\mathrm{edge}} = \varrho_n I_n K_n f_n^2, \quad n \in \mathcal{N}, \tag{8}$$

where ϱ_n is the effective switched capacitance of the edge cloud n. Else, if the task is executed by the central cloud, we then have the central processing energy consumption, including the backhaul transmission and the computation energy consumption, which is expressed as

$$E_n^{\mathrm{central}} = \mathrm{tr}\left(\mathbf{Q}_n\right) T_n^{\mathrm{central}}(\mathbf{Q}) + \zeta_n E_n^{\mathrm{edge}}, \quad n \in \mathcal{N}, \tag{9}$$

where ζ_n is the ratio of the central cloud's computation energy consumption to that of the edge cloud n for computing the same UE n's task.[5] Thus, the total energy consumption for task offloading and computation can be calculated as[6]

$$E_{\mathrm{total}} = \sum_{n=1}^{N} \left(E_n^{\mathrm{a}} + c_n E_n^{\mathrm{edge}} + (1 - c_n) E_n^{\mathrm{central}}\right). \tag{10}$$

[4]In our considered scenario, the accessing latency of MBS to the central cloud through optical fiber should be negligible especially compared with the wireless backhaul transmission latency. For the extreme case that the optical fiber transmission latency is not negligible, the central cloud processing latency can be re-expressed as $T_n^{\mathrm{central}}(\mathbf{Q}) = \frac{I_n}{R_n^{\mathrm{b}}(\mathbf{Q})} + T_{\mathrm{of}}^{\mathrm{central}}$, where $T_{\mathrm{of}}^{\mathrm{central}}$ is a maximum threshold of optical fiber transmission latency. Even though, the proposed algorithms are still effective.

[5]ζ_n can be determined by ϱ_n, f_n, and the effective switched capacitance and the CPU frequency of the central cloud used for computing UE n's task. Different values of $\{\zeta_n, n \in \mathcal{N}\}$ represent different relationships between the computing energy consumption at central cloud and edge clouds, and may have different effects on edge/central cloud selection and system performance.

[6]Here, the static energy consumption of UEs, SBSs, and MBS consumed by the circuit or cooling is ignored since it has negligible effect on our design.

C. Problem Formulation

Our aim is to minimize network's total energy consumption used for task offloading and computation under central/backhaul and edge processing latency constraints through jointly optimizing UEs' cloud selection decisions in $\mathbf{c} = \{c_n\}_{n=1}^N$, UEs' transmit power vector $\mathbf{p}^{\mathrm{u}}$, SBSs' receive beamformers in $\mathbf{w} = \{\mathbf{w}_n\}_{n=1}^N$ and SBSs' transmit covariance matrices in $\mathbf{Q}$. To this end, the problem is formulated as

$$\begin{aligned}
&\min_{\mathbf{c},\mathbf{p}^{\mathrm{u}},\mathbf{w},\mathbf{Q}} \quad E_{\mathrm{total}} \\
&\text{s.t. } \mathrm{C1}: c_n \in \{0,1\}, \quad \forall n \in \mathcal{N}, \\
&\qquad \mathrm{C2}: (1-c_n)\, T_n^{\mathrm{central}}(\mathbf{Q}) \le \alpha T_n^{\mathrm{edge}}, \quad \forall n \in \mathcal{N}, \\
&\qquad \mathrm{C3}: T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) + c_n T_n^{\mathrm{edge}} \le T_{\mathrm{th}}, \quad \forall n \in \mathcal{N}, \\
&\qquad \mathrm{C4}: 0 \le p_n^{\mathrm{u}} \le P_{\max}^{\mathrm{u}}, \quad \forall n \in \mathcal{N}, \\
&\qquad \mathrm{C5}: \mathbf{Q}_n \succeq \mathbf{0}, \quad \forall n \in \mathcal{N}. \qquad (11)
\end{aligned}$$

In problem (11), C2 represents the central/backhaul processing latency constraint, indicating that the central cloud is selected, i.e., the backhaul is allowed to be used for task offloading, only when the setting parameters can make sure that the central/backhaul processing latency is lower than certain percentage, e.g., α, of edge computing latency. Considering the scarce backhaul resources, this constraint is reasonable in practice and of great benefit to guarantee the high-speed backhaul transmission, avoid the abuse of backhaul and alleviate the backhaul congestion. Here, $0 < \alpha < 1$ is a predefined fraction for a specified scenario depending on the central cloud and backhaul restriction. For the special case of $\alpha = 0$, central cloud becomes unavailable as indicated in C2 and thus $c_n = 1$ for $n \in \mathcal{N}$, then problem (11) reduces to resource allocation problem in traditional MEC networks, which has been studied from different perspectives in the literatures such as [12], [18]–[26]. C3 is the latency constraint for edge processing, such that the sum of the access transmission latency and the edge computing latency should not exceed a given threshold T_{th}. Note that T_n^{edge} expressed in (4) increases with the task size I_n, and thus if edge cloud cannot meet its latency constraint in C3 when encounters large tasks, e.g., $T_n^{\mathrm{edge}} > T_{\mathrm{th}}$, central cloud will be the only option to be utilized, which further indicates the complementary relationship between edge and central cloud computing [5]. C4 and C5 guarantee the non-negativeness of the transmit power values.

In our considered scenario, we assume that UEs' tasks have already been synchronized. In fact, our work can be easily extended into the cases considering the latency of synchronizing UEs' tasks. For the case with deterministic task arrival model [4], the edge processing latency constraints C3 should be changed into $T_n^{\mathrm{syn}} + T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) + c_n T_n^{\mathrm{edge}} \le T_{\mathrm{th}}, n \in \mathcal{N}$, where T_n^{syn} is the synchronization latency of UE n. For the case with random task arrival model [4], we can introduce a maximum synchronization latency threshold, denoted as T_{syn}. Then constraints C3 can be changed into $T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) + c_n T_n^{\mathrm{edge}} \le T_{\mathrm{th}} - T_{\mathrm{syn}}, n \in \mathcal{N}$. In this way, we can also leverage the algorithms proposed in section III to solve the corresponding formulated problems.

III. Algorithm Design

The considered problem (11) is a mixed-integer and non-convex optimization problem because of the integer cloud selection indicator $\mathbf{c}$, and the non-convex objective function and constraints C2, C3, which is NP-hard in general and its optimal solution is difficult to achieve. To be tractable, we first need to determine whether edge or central cloud computing will be employed, and then we can optimize the transmit powers, receive beamformers and covariance matrices. Hence, a tractable decomposition approach can be developed to solve (11) in an iterative manner considering the fact that $\mathbf{c}$ and $\{\mathbf{p}^{\mathrm{u}}, \mathbf{w}, \mathbf{Q}\}$ are coupled in the objective function and constraints C2, C3 of problem (11).

A. Edge or Central Cloud Computing

As mentioned in section II-C, when the n-th edge cloud's computing time T_n^{edge} is greater than the maximum allowable time T_{th}, the use of edge cloud is infeasible and central cloud computing has to be utilized, i.e., $c_n = 0$. Next, we optimize the cloud selection indicator $\mathbf{c}$ for the case of $T_n^{\mathrm{edge}} < T_{\mathrm{th}}$, $n \in \mathcal{N}$. To properly deal with the integer optimization caused by c_n, we first relax $c_n \in \{0,1\}$ as $\widehat{c}_n \in [0,1]$, and denote $\widehat{\mathbf{c}} = \{\widehat{c}_n\}_{n=1}^N$ as the set of the relaxed cloud selection variable $\widehat{c}_n$. Then problem (11) with given feasible $\{\mathbf{p}^{\mathrm{u}}, \mathbf{w}, \mathbf{Q}\}$ can be decomposed into the following relaxed version

$$\begin{aligned}
&\min_{\widehat{\mathbf{c}}} \quad \sum_{n=1}^{N} \left(\widehat{c}_n E_n^{\mathrm{edge}} + (1-\widehat{c}_n)\, E_n^{\mathrm{central}}\right) \\
&\text{s.t. } \widehat{\mathrm{C}}1: \widehat{c}_n \in [0,1], \quad \forall n \in \mathcal{N}, \\
&\qquad \widehat{\mathrm{C}}2: (1-\widehat{c}_n)\, T_n^{\mathrm{central}}(\mathbf{Q}) \le \alpha T_n^{\mathrm{edge}}, \quad \forall n \in \mathcal{N}, \\
&\qquad \widehat{\mathrm{C}}3: T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) + \widehat{c}_n T_n^{\mathrm{edge}} \le T_{\mathrm{th}}, \quad \forall n \in \mathcal{N}. \qquad (12)
\end{aligned}$$

Problem (12) is a one-dimensional linear programming, and its solution can be given in the following two cases:

- Case 1: Without loss of generality, if the energy consumption of edge computing is lower than that of central processing for UE n's task, i.e, $E_n^{\mathrm{edge}} \le E_n^{\mathrm{central}}$, the objective function of problem (12) is a decreasing function of $\widehat{c}_n$. Therefore, the optimal $\widehat{c}_n^*$ is the maximum value that satisfies $\widehat{\mathrm{C}}1 - \widehat{\mathrm{C}}3$, i.e.,

$$\widehat{c}_n^* = \left[\min\left\{\frac{T_{\mathrm{th}} - T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n)}{T_n^{\mathrm{edge}}}, 1\right\}\right]^+. \qquad (13)$$

- Case 2: if $E_n^{\mathrm{edge}} > E_n^{\mathrm{central}}$, the objective function of problem (12) is an increasing function of $\widehat{c}_n$, and the optimal $\widehat{c}_n^*$ is the minimum value that satisfies $\widehat{\mathrm{C}}1 - \widehat{\mathrm{C}}3$, i.e.,

$$\widehat{c}_n^* = \left[1 - \frac{\alpha T_n^{\mathrm{edge}}}{T_n^{\mathrm{central}}(\mathbf{Q})}\right]^+. \qquad (14)$$

It is seen that the relaxed edge/central cloud computing decision $\widehat{\mathbf{c}}^*$ is reliant on the optimal $\{\mathbf{p}^{\mathrm{u}}, \mathbf{w}, \mathbf{Q}\}$ of problem (11). In the following two subsections, we will focus on obtaining the optimal $\{\mathbf{p}^{\mathrm{u}*}, \mathbf{w}^*\}$ and $\mathbf{Q}^*$, respectively, based on a given cloud selection decision $\widehat{\mathbf{c}}$.

B. UEs' Transmit Powers and SBSs' Receive Beamformers

For fixed cloud selection decision $\widehat{\mathbf{c}}$, the optimal $\{\mathbf{p}^{\mathrm{u}*}, \mathbf{w}^*\}$ can be obtained by solving a subproblem of (11) as follows:

$$\begin{aligned} &\min_{\mathbf{p}^{\mathrm{u}},\mathbf{w}} \ \sum_{n=1}^{N} p_n^{\mathrm{u}} T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) \\ &\text{s.t. } \widehat{\mathrm{C}}3, \quad \mathrm{C}4, \end{aligned} \tag{15}$$

where $\widehat{\mathrm{C}}3$ and $\mathrm{C}4$ are the corresponding constraints expressed in problem (12) and (11), respectively. The subproblem (15) is non-convex (over $\mathbf{p}^{\mathrm{u}}$) and its objective function is the weighted sum-of-ratios, which is challenging to solve. We first examine the interplay between UEs' transmit power vector $\mathbf{p}^{\mathrm{u}}$ and SBSs' receive beamformers in $\mathbf{w}$.

Lemma 1: For fixed $\mathbf{p}^{\mathrm{u}}$*, the optimal* $\mathbf{w}_n^*$ *of problem* (15) *is given by*

$$\mathbf{w}_n^* = \mathrm{eigvec}\left\{\max\left\{\mathrm{eig}\{(\mathbf{\Omega}_{-n})^{-1}\mathbf{\Omega}_n\}\right\}\right\}, \tag{16}$$

where $\mathbf{\Omega}_{-n} = \sigma_n^2\mathbf{I}_L + \sum_{i=1,i\neq n}^{N} p_i^{\mathrm{u}}\mathbf{h}_{i,n}^{\mathrm{a}}(\mathbf{h}_{i,n}^{\mathrm{a}})^H$ *and* $\mathbf{\Omega}_n = p_n^{\mathrm{u}}\mathbf{h}_{n,n}^{\mathrm{a}}(\mathbf{h}_{n,n}^{\mathrm{a}})^H$.

Proof: See Appendix A. □

With the help of auxiliary variables $\mathbf{t} = \{t_n\}_{n=1}^N$, problem (15) over the UEs' transmit power vector $\mathbf{p}^{\mathrm{u}}$ for fixed $\mathbf{w}$ can be equivalently transformed as

$$\begin{aligned} &\min_{\mathbf{p}^{\mathrm{u}},\mathbf{t}} \ \sum_{n=1}^{N} I_n t_n \\ &\text{s.t. } \widetilde{\mathrm{C}}1: \frac{p_n^{\mathrm{u}}}{R_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n)} \le t_n, \quad \forall n \in \mathcal{N}, \\ &\qquad \widetilde{\mathrm{C}}2: \gamma_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) \ge \tau_n, \quad \forall n \in \mathcal{N}, \\ &\qquad \widetilde{\mathrm{C}}3: 0 \le p_n^{\mathrm{u}} \le P_{\max}^{\mathrm{u}}, \quad \forall n \in \mathcal{N}, \end{aligned} \tag{17}$$

where $\tau_n = 2^{\frac{I_n}{B^{\mathrm{a}}\left(\tau_{\mathrm{th}} - \widehat{c}_n T_n^{\mathrm{edge}}\right)}} - 1$.

Lemma 2: The optimal solution $(\mathbf{p}^{\mathrm{u}*}, \mathbf{t}^*)$ *of problem* (17) *satisfies the Karush-Kuhn-Tucker (KKT) conditions of the following* N $(n \in \mathcal{N})$ *subproblems*

$$\begin{aligned} &\min_{p_n^{\mathrm{u}}} \ \left(\lambda_n + M_n\right) p_n^{\mathrm{u}} - \lambda_n t_n R_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) \\ &\text{s.t. } \widetilde{\mathrm{C}}2: \gamma_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n) \ge \tau_n, \\ &\qquad \widetilde{\mathrm{C}}3: 0 \le p_n^{\mathrm{u}} \le P_{\max}^{\mathrm{u}}, \end{aligned} \tag{18}$$

with

$$M_n = \sum_{j=1,j\neq n}^{N} \lambda_j t_j \frac{B_{\mathrm{a}}}{\ln 2} \frac{\left(\gamma_j^{\mathrm{a}}\right)^2 |\mathbf{w}_j^H \mathbf{h}_{n,j}^{\mathrm{a}}|^2}{p_j^{\mathrm{u}} |\mathbf{w}_j^H \mathbf{h}_{j,j}^{\mathrm{a}}|^2 \left(1+\gamma_j^{\mathrm{a}}\right)} + \sum_{j=1,j\neq n}^{N} \mu_j \frac{\left(\gamma_j^{\mathrm{a}}\right)^2 |\mathbf{w}_j^H \mathbf{h}_{n,j}^{\mathrm{a}}|^2}{p_j^{\mathrm{u}} |\mathbf{w}_j^H \mathbf{h}_{j,j}^{\mathrm{a}}|^2}, \tag{19}$$

where $\{\lambda_n\}_{n=1}^N$ *and* $\{\mu_n\}_{n=1}^N$ *are Lagrange multipliers associated with the constraints* $\widetilde{\mathrm{C}}1$ *and* $\widetilde{\mathrm{C}}2$ *of problem* (17)*, respectively, and* $M_n = -\sum_{j\neq n}^{N} \lambda_j t_j \frac{\partial R_j^{\mathrm{a}}}{\partial p_n^{\mathrm{u}}} - \sum_{j\neq n}^{N} \mu_j \frac{\partial \gamma_j^{\mathrm{a}}}{\partial p_n^{\mathrm{u}}}$.

For optimal $(\mathbf{p}^{\mathrm{u}*}, \mathbf{w}^*)$*,* λ_n *and* t_n *are respectively calculated as*

$$\lambda_n = \frac{I_n}{R_n^{\mathrm{a}}\left(\mathbf{p}^{\mathrm{u}*}, \mathbf{w}_n^*\right)}, \tag{20}$$

$$t_n = \frac{p_n^{\mathrm{u}*}}{R_n^{\mathrm{a}}\left(\mathbf{p}^{\mathrm{u}*}, \mathbf{w}_n^*\right)}. \tag{21}$$

Proof: See Appendix B. □

Given λ_n and t_n, the subproblem (18) is convex w.r.t. p_n^{u}. Therefore, we have the following theorem.

Theorem 1: The solution of subproblem (18) *is given by*

$$p_n^{\mathrm{u}*} = \begin{cases} \dfrac{\tau_n}{\Lambda_n}, & \text{if } G_n < \dfrac{\tau_n}{\Lambda_n}, \\ G_n, & \text{if } \dfrac{\tau_n}{\Lambda_n} \le G_n \le P_{\max}^{\mathrm{u}}, \\ P_{\max}^{\mathrm{u}}, & \text{if } G_n > P_{\max}^{\mathrm{u}}, \end{cases} \tag{22}$$

$$\mu_n^* = \begin{cases} \dfrac{\lambda_n + M_n}{\Lambda_n} - \dfrac{B_{\mathrm{a}}}{\ln 2}\dfrac{\lambda_n t_n}{\tau_n + 1}, & \text{if } G_n < \dfrac{\tau_n}{\Lambda_n}, \\ 0, & \text{otherwise}, \end{cases} \tag{23}$$

$$\nu_n^* = \begin{cases} \dfrac{B_{\mathrm{a}}}{\ln 2}\dfrac{\lambda_n t_n}{P_{\max}^{\mathrm{u}} + 1/\Lambda_n} - \lambda_n - M_n, & \text{if } G_n > P_{\max}^{\mathrm{u}}, \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

where we define $\Lambda_n \triangleq \frac{|\mathbf{w}_n^H \mathbf{h}_{n,n}^{\mathrm{a}}|^2}{\sum_{i=1,i\neq n}^{N} p_i^{\mathrm{u}} |\mathbf{w}_n^H \mathbf{h}_{i,n}^{\mathrm{a}}|^2 + |\mathbf{w}_n^H \mathbf{n}_n|^2}$*,* $G_n \triangleq \frac{B_{\mathrm{a}}}{\ln 2}\frac{\lambda_n t_n}{\lambda_n + M_n} - \frac{1}{\Lambda_n}$*, and* μ_n^* *and* ν_n^* *are respectively the optimal Lagrange multipliers associated with the constraints* $\widetilde{\mathrm{C}}2$ *and* $\widetilde{\mathrm{C}}3$ *of problem* (18)*.*

Proof: See Appendix C. □

In light of the results in **Lemma** 1, **Lemma** 2 and **Theorem** 1, we provide an iterative approach to effectively solve problem (15), which is shown in **Algorithm** 1.

Algorithm 1 Solution of Problem (15)

1: **Initialize** $p_n^{\mathrm{u}} = P_{\max}^{\mathrm{u}}$, $\forall n$. Set $\mathbf{w}_n$ based on **Lemma** 1.
2: **Repeat**
3: a) Given $\mathbf{w}$, Loop:
i): Compute M_n, λ_n and t_n based on **Lemma** 2.
ii): Update p_n^{u} and μ_n based on **Theorem** 1.
Until convergence.
4: b) Update $\mathbf{w}$ based on **Lemma** 1.
5: Until convergence, and obtain the optimal $\{\mathbf{p}^{\mathrm{u}*}, \mathbf{w}^*\}$.

The convergence of **Algorithm** 1 can be guaranteed since the objective function of problem (15) decreases with the iteration index (in step 3 and step 4 of **Algorithm** 1), which is indicated from optimizing $\mathbf{p}^{\mathrm{u}}$ and $\mathbf{w}$ in each iteration as shown in **Lemma** 1 and **Lemma** 2, respectively.

C. SBSs' Transmit Covariance Matrices

For fixed cloud selection decision $\widehat{\mathbf{c}}$, the optimal $\mathbf{Q}^*$ can be obtained by solving the following subproblem:

$$\begin{aligned} &\min_{\mathbf{Q}} \ y(\mathbf{Q}) = \sum_{n=1}^{N} \left(1 - \widehat{c}_n\right) \mathrm{tr}\left(\mathbf{Q}_n\right) T_n^{\mathrm{central}}(\mathbf{Q}) \\ &\text{s.t. } \widehat{\mathrm{C}}2: R_n^{\mathrm{b}}(\mathbf{Q}) \ge \left(1 - \widehat{c}_n\right) \frac{I_n}{\alpha T_n^{\mathrm{edge}}}, \quad \forall n \in \mathcal{N}, \quad \mathrm{C}5, \end{aligned} \tag{25}$$

where $\widehat{\mathrm{C}}2$ and C5 are the corresponding constraints expressed in problem (12) and (11), respectively, and $\widehat{\mathrm{C}}2$ is re-expressed in an equivalent form here. Problem (25) is non-convex due to the non-convexity of the objective function and constraint $\widehat{\mathrm{C}}2$, which cannot be solved directly. Thus, we resort to a successive pseudoconvex approach, which has many advantages such as fast convergence and parallel computation [38].

First, let $\mathbf{Q}^l$ denote the $\mathbf{Q}$ value in the l-th iteration. Thus the non-convex item $(1-\widehat{c}_n)\mathrm{tr}\left(\mathbf{Q}_n\right)T_n^{\mathrm{central}}(\mathbf{Q})$ for each $n\in\mathcal{N}$ in the objective function can be approximated as a pseudoconvex function at $\mathbf{Q}^l$, which is written as

$$\widehat{y}_n(\mathbf{Q}_n;\mathbf{Q}^l)\triangleq(1-\widehat{c}_n)\frac{I_n\mathrm{tr}(\mathbf{Q}_n)}{R_n^{\mathrm{b}}(\mathbf{Q}_n;\mathbf{Q}^l)}+\chi_n(\mathbf{Q}_n),\quad(26)$$

where $\chi_n(\mathbf{Q}_n)=\sum_{j\neq n}I_j\mathrm{tr}(\mathbf{Q}_j^l)\left\langle(\mathbf{Q}_n-\mathbf{Q}_n^l),\nabla_{\mathbf{Q}_n^\dagger}\frac{1-\widehat{c}_j}{R_j^{\mathrm{b}}(\mathbf{Q}^l)}\right\rangle$ is a function obtained by linearizing the non-convex function $\sum_{j\neq n}^{N}(1-\widehat{c}_j)\mathrm{tr}\left(\mathbf{Q}_j\right)T_j^{\mathrm{central}}(\mathbf{Q})$ in $\mathbf{Q}_n$ at the point $\mathbf{Q}^l$. Based on (26), we can approximate the objective function $y\left(\mathbf{Q}\right)$ of problem (25) at $\mathbf{Q}^l$ as

$$\widehat{y}(\mathbf{Q};\mathbf{Q}^l)=\sum_{n=1}^{N}\widehat{y}_n(\mathbf{Q}_n;\mathbf{Q}^l).\quad(27)$$

It is easily seen that $\widehat{y}(\mathbf{Q};\mathbf{Q}^l)$ is pseudoconvex and has the same gradient with $y\left(\mathbf{Q}\right)$ at $\mathbf{Q}=\mathbf{Q}^l$ [38].

Then, by equivalently rewriting the non-concave function $R_n^{\mathrm{b}}(\mathbf{Q})$ in constraint $\widehat{\mathrm{C}}2$ as a difference of two concave functions as expressed in (28a) according to its definition in (6), and leveraging the first-order Taylor expansion at $\mathbf{Q}^l$ for the second concave function denoted as $R_n^{\mathrm{b}2}(\mathbf{Q})=B^{\mathrm{b}}\log_2\det\left(\sigma^2\mathbf{I}+\sum_{i\neq n}^{N}\mathbf{H}_i^{\mathrm{b}}\mathbf{Q}_i\left(\mathbf{H}_i^{\mathrm{b}}\right)^H\right)$, $R_n^{\mathrm{b}}(\mathbf{Q})$ can be approximated as

$$\begin{aligned}R_n^{\mathrm{b}}(\mathbf{Q})&=B^{\mathrm{b}}\log_2\det\left(\sigma^2\mathbf{I}+\boldsymbol{\Xi}(\mathbf{Q})\right)-R_n^{\mathrm{b}2}(\mathbf{Q})\quad(28a)\\&\geqslant B^{\mathrm{b}}\log_2\det\left(\sigma^2\mathbf{I}+\boldsymbol{\Xi}(\mathbf{Q})\right)-R_n^{\mathrm{b}2}(\mathbf{Q}^l)\\&\quad-\sum_{j\neq n}^{N}\left\langle(\mathbf{Q}_j-\mathbf{Q}_j^l),\nabla_{\mathbf{Q}_j^\dagger}R_n^{\mathrm{b}2}(\mathbf{Q}^l)\right\rangle\triangleq\bar{R}_n^{\mathrm{b}}(\mathbf{Q}),\quad(28b)\end{aligned}$$

where $\boldsymbol{\Xi}\left(\mathbf{Q}\right)=\sum_{i=1}^{N}\mathbf{H}_i^{\mathrm{b}}\mathbf{Q}_i\left(\mathbf{H}_i^{\mathrm{b}}\right)^H$. Here, $\bar{R}_n^{\mathrm{b}}(\mathbf{Q})$ expressed in (28b) is a concave function over $\mathbf{Q}$.

Therefore, at $\mathbf{Q}^l$, the original problem (25) can be approximately transformed as

$$\begin{aligned}&\min_{\mathbf{Q}}\ \widehat{y}(\mathbf{Q};\mathbf{Q}^l)\\&\text{s.t. }\overline{\mathrm{C}}2:\bar{R}_n^{\mathrm{b}}(\mathbf{Q})\geq(1-\widehat{c}_n)\frac{I_n}{\alpha T_n^{\mathrm{edge}}},\quad\forall n\in\mathcal{N},\quad\mathrm{C5}.\quad(29)\end{aligned}$$

The objective function of problem (29) is a sum of N pseudoconvex functions each containing a fractional function and a linear function. In addition, all the constraints in problem (29) are convex. Hence, by leveraging the Dinkelbach-like algorithm [39] and introducing a set of auxiliary variables for the N fractional functions in the objective function, problem (29) can be transformed into a solvable convex optimization problem, which can be effectively solved by CVX [40] and owns provable convergence [38]. Let $\mathbb{B}\mathbf{Q}^l$ represent the optimal solution of problem (29) at the l-th iteration, and thus the value of $\mathbf{Q}$ in the next $(l+1)$-th iteration can be updated as

$$\mathbf{Q}^{l+1}=\mathbf{Q}^l+\varsigma(l)(\mathbb{B}\mathbf{Q}^l-\mathbf{Q}^l),\quad(30)$$

where $\varsigma(l)$ is the step size at the l-th iteration and can be obtained through the successive line search, and $\mathbb{B}\mathbf{Q}^l-\mathbf{Q}^l$ is the descent direction of $y\left(\mathbf{Q}\right)$. Thus, the solution of problem (25) can be iteratively obtained.

Based on the aforementioned analysis of optimizing the variables $\{\mathbf{p}^{\mathrm{u}*},\mathbf{w}^*,\mathbf{Q}^*\}$, **Algorithm** 2 is proposed to solve the original problem (11).

Algorithm 2 Solution of Problem (11)

1: **Initialize** $p_n^{\mathrm{u}}=P_{\max}^{\mathrm{u}}$, $\forall n$. Set $\mathbf{w}_n$ based on **Lemma** 1.
Based on the constraint $\widehat{\mathrm{C}}3$ of problem (12), set $\widehat{c}_n=\left[\min\left\{\frac{T_{\mathrm{th}}-T_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}},\mathbf{w}_n)}{T_n^{\mathrm{edge}}},1-\delta\right\}\right]^+$, where $\delta\in(0,0.5)$ is a tolerant value to avoid the selection of solely edge clouds or central cloud at the initial point. Then, based on the constraint $\widehat{\mathrm{C}}2$ of problem (12), $\mathbf{Q}$ is set to meet $T_n^{\mathrm{central}}(\mathbf{Q})=\frac{\alpha T_n^{\mathrm{edge}}}{1-\widehat{c}_n}$ through the use of ZF precoding with equal power allocation at each SBS.
2: **Repeat**
3: a) Given $\{\widehat{c}_n\}_{n=1}^N$:
i): Update $\{\mathbf{p}^{\mathrm{u}},\mathbf{w}\}$ based on **Algorithm** 1.
ii): Loop:
ii-1): Solve problem (29) via Dinkelbach-like algorithm [39].
ii-2): Update $\mathbf{Q}^l$ based on (30).
Until convergence, and obtain the updated $\mathbf{Q}$.
4: b) Update $\{\widehat{c}_n\}_{n=1}^N$ according to subsection III-A.
5: Until convergence, and obtain solution $\{\mathbf{c}^*,\mathbf{p}^{\mathrm{u}*},\mathbf{w}^*,\mathbf{Q}^*\}$, in which $\mathbf{c}^*$ is obtained by rounding the cloud selection solution of problem (12), i.e., $\widehat{\mathbf{c}}$, and $\mathbf{p}^{\mathrm{u}*},\mathbf{w}^*,\mathbf{Q}^*$ are obtained based on the final obtained $\mathbf{c}^*$.

D. Convergence and Complexity

The convergence of **Algorithm** 2 is easy to prove in light of the guaranteed convergence of **Algorithm** 1, the Dinkelbach-like algorithm used to solve problem (29) [39], and the update process of the cloud selection $\widehat{\mathbf{c}}$ illustrated in Section III-A. Note that the objective function of problem (11), i.e., the network's total energy consumption for task offloading and computation, is a decreasing function of the iteration index (in step 3 and step 4 of **Algorithm** 2), which ensures the convergence of **Algorithm** 2.

The proposed **Algorithm** 2 enjoys an acceptable complexity as well as an easy implementation. In each iteration, the majority of computational complexity lies in solving subproblem (15) for obtaining the optimal $(\mathbf{p}^{\mathrm{u}*},\mathbf{w}^*)$ and the approximate subproblem (29) for obtaining the optimal $\mathbf{Q}^*$ with a given $\widehat{\mathbf{c}}$. In the proposed algorithm, problem (15) can be equivalently transformed into N independent subproblems (18) and thus can be easily solved in a parallel way. Moreover, the optimal solution of each subproblem has

closed-form expressions as indicated in **Theorem** 1, which only generates a complexity ordered by $\mathcal{O}(N)$. For the approximate subproblem (29), the Dinkelbach-like algorithm is proved to exhibit a linear convergence rate [39] and the corresponding convex optimization problem can be efficiently solved by CVX, thus the generated complexity is acceptable in general.

In order to further reduce the complexity of solving the optimization problem for minimizing the network's total energy consumption of task offloading and computation, we will consider the case of applying the massive MIMO technology at the MBS in the following section. It demonstrates that the complexity of the proposed algorithm can be substantially reduced while even better performance can be achieved compared to the case with traditional MIMO backhaul.

IV. Massive MIMO Backhaul

In the prior sections, we have studied the synergy of combining edge-central cloud computing with traditional multi-cell MIMO backhaul. Since massive MIMO has been one of the key 5G radio-access technologies, in this section, we further consider the time-division duplex (TDD) massive MIMO aided backhaul in the Rayleigh fading environment, i.e., the MBS is equipped with a very large number of antennas and the SBSs only use one single transmit antenna ($M \gg N$).

There are two main merits for massive MIMO backhaul transmissions: 1) Since SBSs and MBSs are usually still and the backhaul channels will become deterministic, a phenomenon known as "channel hardening" [41], [42], and thus the backhaul channel coherence time will be much longer than ever before, which means that the time spent on uplink channel estimation will be much lower. 2) As shown in [43], simple linear processing methods can achieve nearly-optimal performance. As a result, we will consider two linear detection schemes at the MBS, namely MRC and ZF, to provide low-complexity massive MIMO backhaul solutions.

A. MRC Receiver at the MBS

When MRC receiver is applied at the MBS, we consider a lower-bound achievable backhaul rate for tractability, which can well approximate the exact massive MIMO transmission rate as confirmed in [44]. As such, given the cloud selection decision $\widehat{\mathbf{c}}$, the backhaul related problem (25) reduces to

$$\begin{aligned} \min_{\mathbf{q}} \ & \sum_{n=1}^{N} (1-\widehat{c}_n)\, q_n \frac{I_n}{R_n^{\mathrm{b}}(\mathbf{q})} \\ \text{s.t. } & \widehat{\mathrm{C}}2: R_n^{\mathrm{b}}(\mathbf{q}) \geq (1-\widehat{c}_n)\, \frac{I_n}{\alpha T_n^{\mathrm{edge}}}, \quad \forall n \in \mathcal{N}, \\ & \mathrm{C}5: q_n \geq 0, \quad \forall n \in \mathcal{N}, \end{aligned} \tag{31}$$

where q_n is the n-th SBS's transmit power, $\mathbf{q} = [q_1, \cdots, q_N]$, and $R_n^{\mathrm{b}}(\mathbf{q}) = B^{\mathrm{b}} \log_2 \left(1 + (M-1) \frac{q_n \beta_n}{\sum_{i=1, i\neq n}^{N} q_i \beta_i + \sigma_n^2}\right)$, in which β_i is the large-scale fading coefficient of the link between SBS i and the MBS [44]. Problem (31) is non-convex, but can be equivalent to problem (15) with $\mathbf{w}_n = \mathbf{1}$. Thus, it can be directly solved by using **Algorithm** 1. Note that when using **Algorithm** 1, SBSs' initial feasible transmit power vector $\mathbf{q}$ needs to be carefully selected. Here, we assume that the present fractional power control solution applied in 3GPP-LTE [45] can satisfy the constraint $\widehat{\mathrm{C}}2$ in (31), i.e., $q_n = (d_n)^{\epsilon \varpi^{\mathrm{b}}}$, where d_n is the communication distance between the n-th SBS and the MBS, $\epsilon \in [0, 1]$ is the pathloss compensation factor, and ϖ^{b} is the pathloss exponent of the backhaul link. For the special case of full compensation ($\epsilon = 1$), the number of MBS's antennas needs to meet

$$M \geq 1 + (N-1)\left(2^{\frac{(1-\widehat{c}_n) I_n}{B^{\mathrm{b}} \alpha T_n^{\mathrm{edge}}}} - 1\right). \tag{32}$$

B. ZF Receiver at the MBS

When ZF receiver is applied at the MBS, we adopt the corresponding tight lower-bound achievable rate shown in [44]. Given the cloud selection decision $\widehat{\mathbf{c}}$, the backhaul related problem (25) reduces to the following version

$$\begin{aligned} \min_{\mathbf{q}} \ & \sum_{n=1}^{N} (1-\widehat{c}_n)\, \frac{q_n I_n}{R_n^{\mathrm{b}}(q_n)} \\ \text{s.t. } & \widehat{\mathrm{C}}2: R_n^{\mathrm{b}}(q_n) \geq (1-\widehat{c}_n)\, \frac{I_n}{\alpha T_n^{\mathrm{edge}}}, \quad \forall n \in \mathcal{N}, \\ & \mathrm{C}5: q_n \geq 0, \quad \forall n \in \mathcal{N}, \end{aligned} \tag{33}$$

where $R_n^{\mathrm{b}}(q_n) = B^{\mathrm{b}} \log_2 \left(1 + (M-N) \frac{q_n \beta_n}{\sigma_n^2}\right)$. Since $\frac{q_n}{R_n^{\mathrm{b}}(q_n)}$ is an increasing function of q_n ($\frac{\partial\left(\frac{q_n}{R_n^{\mathrm{b}}(q_n)}\right)}{\partial q_n} \geq 0$), the optimal q_n^* is the minimum value that meets the constraints $\widehat{\mathrm{C}}2$ and $\mathrm{C}5$ in (33), i.e.,

$$q_n^* = \frac{2^{\frac{(1-\widehat{c}_n) I_n}{B^{\mathrm{b}} \alpha T_n^{\mathrm{edge}}}} - 1}{(M-N)\frac{\beta_n}{\sigma_n^2}}, \quad \forall n \in \mathcal{N}. \tag{34}$$

Based on the above analysis, when massive MIMO backhaul is employed at the MBS, the solution of problem (11) can still be obtained by using the proposed **Algorithm** 2, where the optimal SBSs' transmit powers are given by the solution of problem (31) for the MRC receiver or (34) for the ZF receiver.

In comparison with the case of using traditional MIMO backhaul, the MRC and ZF linear detection schemes for the case with massive MIMO backhaul links can enjoy super-low complexity. For MRC scheme, the problem (31) can be effectively solved by **Algorithm** 1, and its computational complexity is with the order of $\mathcal{O}(N)$. For ZF scheme, the closed-form solution of problem (33) can be directly obtained, and its complexity order is $\mathcal{O}(1)$. Hence, applying the massive MIMO technology at the MBS can significantly facilitate the cooperation between the edge and central clouds by providing easier but more efficient backhaul offloading for UEs to access the central cloud computing services.

V. Simulation Results

In this section, simulation results are presented to evaluate the performance of the proposed algorithms and shed light on the effects of the key parameters including the ratio of energy consumption between central and edge cloud computing

TABLE I
SIMULATION PARAMETERS

Parameter	Symbol	Value
Bandwidth for an access or backhaul link	B^{a},B^{b}	10 MHz
Noise power spectral density for an access or backhaul link	$\sigma_n^2, n \in \mathcal{N},\ \sigma^2$	-174 dBm/Hz
Pathloss exponent for access links	ϖ^{a}	3.67
Pathloss exponent for backhaul links	ϖ^{b}	2.35
Pathloss compensation factor	ϵ	1
Radius of the small cells	r^{a}	50 m
Radius of the macro cell	r^{b}	500 m
Number of SBSs/UEs	N	6
Number of antennas for each SBS	L	2
UEs' maximum transmit power	$P_{\max}^{\mathrm{u}}$	23 dBm
Required CPU cycles per bit	$K_n,\ n \in \mathcal{N}$	300 cycles/bit
the effective switched capacitance of the SBSs' processors	$\varrho_n,\ n \in \mathcal{N}$	10^{-28}
The tolerant value in **Algorithm** 2	δ	0.1

$(\zeta_n = \zeta, n \in \mathcal{N})$, the task size $(I_n = I, n \in \mathcal{N})$, the latency threshold of edge processing (T_{th}), the required fraction of edge computing time for backhaul transmission (α), and the edge clouds' CPU clock frequency $(f_n = f, n \in \mathcal{N})$. The performance of some practical schemes are also given as benchmarks, including the "Edge-cloud-only", "Central-cloud-only" schemes, and a scheme with fixed cloud selection, denoted as "Half edge, Half central" scheme where half number of UEs choose edge clouds and the other half use central cloud to complete their computation tasks. Besides, the "Initial feasible solution", representing the case with the initial values setting in **Algorithm** 2, is also given as a baseline to show the performance improvement of optimizing some system parameters. Note that the performance indicators (the total energy consumption and the percentage of UEs that select edge cloud computing) shown in the following figures are averaged over 500 independent channel realizations. All the small-scale fading channel coefficients follow independent and identically complex Gaussian distribution with zero mean and unit variance. The pathloss between SBSs and UEs and between MBS and SBSs are respectively set as $-(140.7 + 36.7\log_{10} d)$dB and $-(100.7 + 23.5\log_{10} d)$dB according to 3GPP TR 36.814 [46], where d (in kilometer) is the distance between two nodes. The other basic simulation parameters are listed in Table I.

A. Improvement With Traditional MIMO Backhaul

In this subsection, numerical results for the integrated edge and central cloud computing system with traditional MIMO backhaul are presented in comparison with the benchmarks mentioned before. These results can properly demonstrate the performance enhancement of using the proposed algorithm through jointly optimizing the key system parameters including cloud selection decision, UEs' transmit powers, SBSs' receive beamformers and transmit covariance matrices.

Fig. 2 shows the effect of the uniform computing energy ratio $\zeta = \zeta_n, n \in \mathcal{N}$ on the total energy consumption of the system with traditional MIMO backhaul. We see that the energy consumption of all the schemes are non-decreasing functions of ζ, due to the fact that the energy cost of central cloud computing increases with ζ. It is confirmed that the proposed solution outperforms all the baselines, i.e., the

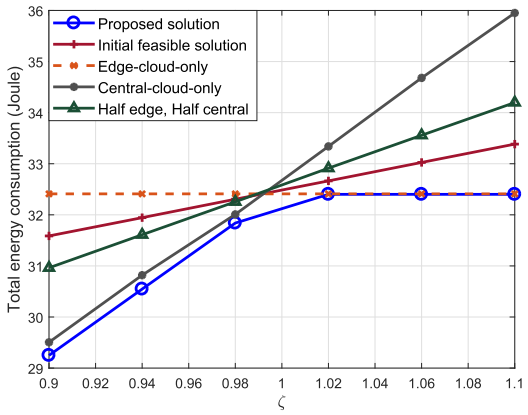


Fig. 2. The total energy consumption of the system with traditional MIMO backhaul versus the uniform computing energy ratio ζ: $M = 16$, $T_{\mathrm{th}} = 0.3$ s, $\alpha = 0.1$, $I = I_n = 5$ Mbits, $f = f_n = 6$ GHz for $n \in \mathcal{N}$.

energy cost can be significantly reduced. The performance improvement is particularly noticeable compared with the Edge-cloud-only scheme in the range of $\zeta < 1$, the traditional Central-cloud-only scheme in the range of $\zeta > 1$, and the Half edge, Half central scheme in the whole range of ζ. In addition, the proposed solution also consumes much less energy than the Initial feasible solution, demonstrating the performance enhancement of jointly optimizing the system parameters.

Fig. 3 depicts the total energy consumption of the system versus the uniform task sizes $I = I_n, n \in \mathcal{N}$ for the cases of $\zeta = 0.9$ and $\zeta = 1.1$. It is easy to understand that computing more input data consumes more energy, and thus the energy cost of each scheme increases with I. Again, we see that the proposed solution is superior to the baseline solutions in all the cases. For the case of $\zeta = 0.9$, the performance of the Central-cloud-only solution is very close to the proposed one since central cloud is dominant in this case, i.e., more UEs tend to use central cloud computing for saving energy. For the case of $\zeta = 1.1$, the advantage of the proposed scheme becomes more obvious compared with the baselines, and actually this case is more common in practice since the central cloud tends to consume more energy for computing because of the higher CPU frequency. We observe that the results of the proposed solution approach to those of the Central-cloud-only solution when I becomes large, indicating that more UEs tend to select

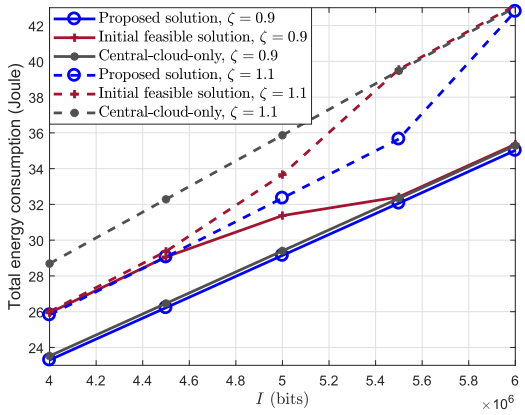


Fig. 3. The total energy consumption of the system with traditional MIMO backhaul versus the uniform task size I: $M = 16$, $T_{\rm th} = 0.3$ s, $\alpha = 0.1$, $f = f_n = 6$ GHz for $n \in \mathcal{N}$.

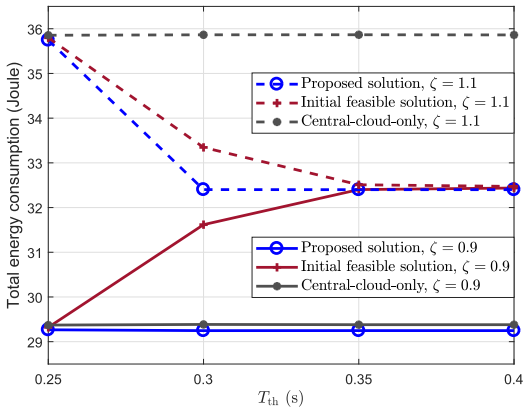


Fig. 4. The total energy consumption of the system with traditional MIMO backhaul versus the latency threshold of edge processing $T_{\rm th}$: $M = 16$, $\alpha = 0.1$, $I = I_n = 5$ Mbits, $f = f_n = 6$ GHz for $n \in \mathcal{N}$.

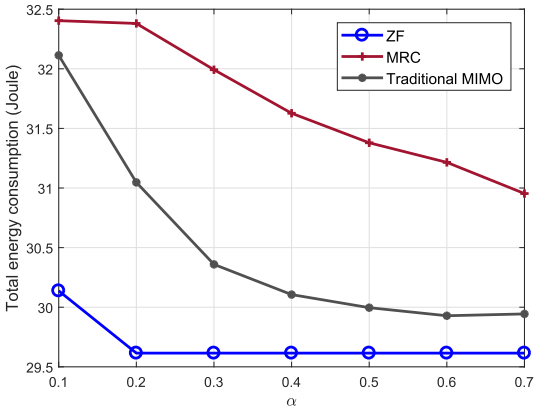


Fig. 5. The total energy consumption of the system versus α: $M = 128$ for massive MIMO backhaul, $M = 8$ for traditional MIMO backhaul, $T_{\rm th} = 0.3$ s, $\zeta = \zeta_n = 0.9$, $I = I_n = 5$ Mbits, $f = f_n = 6$ GHz for $n \in \mathcal{N}$.

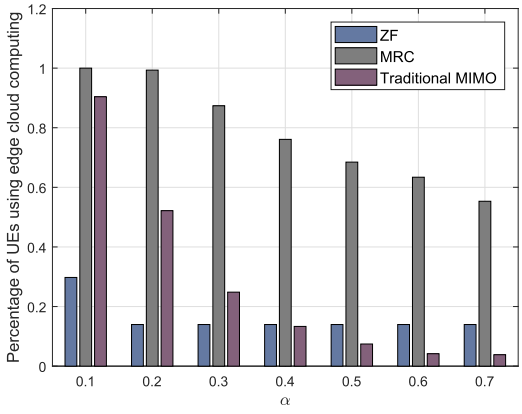


Fig. 6. The percentage of UEs that select edge cloud computing versus α: $M = 128$ for massive MIMO backhaul, $M = 8$ for traditional MIMO backhaul, $T_{\rm th} = 0.3$ s, $\zeta = \zeta_n = 0.9$, $I = I_n = 5$ Mbits, $f = f_n = 6$ GHz for $n \in \mathcal{N}$.

the central cloud for computing, i.e., central cloud computing plays an important role in dealing with relatively large tasks. The reason is that when the task size is large, the edge processing latency constraint C3 of problem (11) may be no longer satisfied due to the limited edge computing capability, and central cloud has to be chosen for computation.

Fig. 4 shows the total energy consumption of the system varying with the latency threshold of edge processing for the cases of $\zeta = 0.9$ and $\zeta = 1.1$. It is seen that the proposed solution is a non-increasing function of $T_{\rm th}$ and outperforms the baselines in both cases. The Central-cloud-only solution is insensitive to $T_{\rm th}$, and its performance is almost invariant thanks to its super computing capability for low computing latency. Note that all the solutions consume almost same amount of energy when $T_{\rm th}$ is small, e.g., $T_{\rm th} = 0.25$ s in this figure. The reason is that the edge processing latency constraint C3 cannot be met and only central cloud computing can be employed to satisfy the latency constraints. For the case of $\zeta = 0.9$, the performance gap between the proposed solution and the Central-cloud-only is small since central cloud computing is dominant, and both solutions perform better than the Initial feasible solution. It is interesting to note that the Initial feasible solution is an increasing function of $T_{\rm th} \in [0.25, 0.4]$ s when $\zeta = 0.9$. This is because the edge cloud computing becomes more feasible as $T_{\rm th}$ increases, and the initial solution allowing more UEs to choose edge cloud for computing while in fact central cloud computing saves more energy, which indicates the importance of optimizing cloud selection in improving the system performance. For the case of $\zeta = 1.1$, the consumed energy of the proposed solution decreases with $T_{\rm th}$ since more UEs are allowed to choose the energy-efficient edge cloud computing for large $T_{\rm th}$.

B. Benefits of Massive MIMO Backhaul

In this subsection, we mainly illustrate the performance of the considered heterogeneous edge/central cloud computing system with massive MIMO backhaul, to confirm the benefits of equipping massive antennas at the MBS in improving the system performance. Here, we focus on MRC and ZF beamforming at the MBS, as studied in Section IV.

Fig. 5 and Fig. 6 depict the total energy consumption and the corresponding percentage of UEs that select edge cloud for computing versus α, respectively. It is seen from Fig. 5 that the energy consumption of each scheme decreases with α since less power will be consumed for backhaul transmission

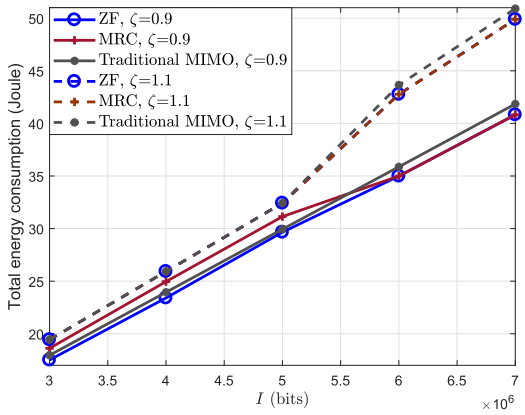
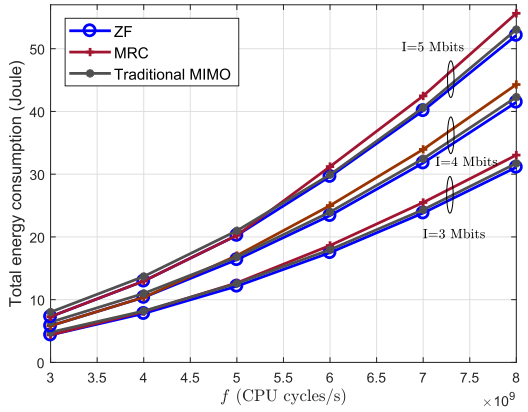
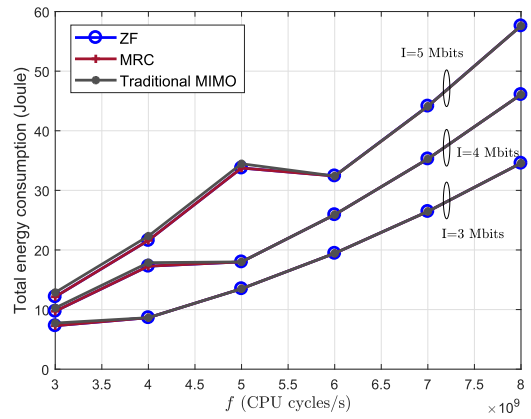


Fig. 7. The total energy consumption of the system versus the uniform task size I: $M = 128$ for massive MIMO backhaul, $M = 8$ for traditional MIMO backhaul, $T_{\rm th} = 0.3$ s, $\alpha = 0.6$, $f = f_n = 6$ GHz for $n \in \mathcal{N}$.



(a) $M = 128$ for massive MIMO backhaul, $M = 8$ for traditio MIMO backhaul, $T_{\rm th} = 0.3$ s, $\alpha = 0.6$, $\zeta = \zeta_n = 0.9$ for $n \in \mathcal{N}$.



(b) $M = 128$ for massive MIMO backhaul, $M = 8$ for traditional MIMO backhaul, $T_{\rm th} = 0.3$ s, $\alpha = 0.6$, $\zeta = \zeta_n = 1.5$ for $n \in \mathcal{N}$.

Fig. 8. The total energy consumption of the system versus SBSs' uniform CPU clock frequency f.

with a higher α according to the backhaul latency constraint C2 of problem (11). This result is also reflected by Fig. 6 where the percentage of UEs using edge cloud computing decreases, which means that more UEs choose to use central cloud for computing as α increases so as to save more energy. Obviously, the energy consumed by the ZF scheme is less than that of the MRC scheme and the solution with traditional MIMO backhaul, which demonstrates the benefits of using ZF beamforming and large antenna arrays at the MBS. Moreover, for the ZF scheme, the percentage of UEs using edge clouds is lower than that of the MRC and traditional MIMO schemes when $\alpha < 0.4$. In contrast, the MRC scheme only uses the edge clouds for computing when $\alpha \leq 0.2$. This is because the backhaul latency constraint C2 in (11) for central cloud processing cannot be satisfied with a small α when MRC receiver is adopted at the MBS due to the inter-SBS interference. Based on these two figures, we see that the consumed energy of the ZF scheme as well as the corresponding percentage of UEs served by edge clouds decrease very slowly, and is almost unchanged for $\alpha \geq 0.2$, which further indicates that the ZF scheme can provide more stable and higher-speed backhaul transmission for computation tasks offloading.

Fig. 7 shows the total energy consumption of the system versus the uniform task size I for the cases of $\zeta = 0.9$ and $\zeta = 1.1$. Similar to Fig. 3, all the curves increase with I as expected. The ZF scheme outperforms the MRC scheme and the traditional MIMO scheme. For the case of $\zeta = 0.9$, the ZF scheme and the traditional MIMO scheme are dominated by central cloud computing, while the MRC scheme experiences a gradual transition from edge-cloud-dominant to central-cloud-dominant and more UEs choose to use central cloud for computing so as to satisfy the processing latency constraint as well as saving energy. For the case of $\zeta = 1.1$, all the schemes are edge-cloud dominant when $I \leq 5$ Mbits, and then gradually become central-cloud-dominant as I increases. It is confirmed that the ZF scheme with massive MIMO backhaul has the advantage of handling the computation-intensive tasks.

Fig. 8(a) and Fig. 8(b) depict the total energy consumption of the system versus the edge clouds' uniform CPU clock frequency $f = f_n, n \in \mathcal{N}$ in the cases of $\zeta = 0.9$ and $\zeta = 1.5$, respectively. According to these two figures, we see that the effect of f is heavily reliant on both the computing task size I and ζ. When I is not large and $\zeta < 1$, network's energy consumption may increase with f as shown in Fig. 8(a), where the curves of all the schemes increase with f and the increasing rates become higher when enlarging I. This is due to the fact that when I is not large and $\zeta < 1$, the energy consumption of the central cloud computing plays a dominant role in contributing to the total energy consumption. In this case, the advantage of using ZF scheme becomes more obvious as f grows large. However, when $\zeta > 1$, network's energy consumption may decrease with f in certain scenario as shown in Fig. 8(b), where there is an obvious decrease as $f \in [5, 6] \times 10^9$ cycles/s in the case of $I = 5$ Mbits. The reason is that when f is small, e.g., less than 4×10^9 cycles/s in Fig. 8(b), the edge processing latency constraint C3 may not be satisfied and central cloud computing becomes the only option. As f increases, edge cloud computing becomes feasible for more UEs to save energy, and the total energy cost will decrease accordingly. In addition, it is seen from Fig. 8(b) that the energy consumption of the three considered schemes are very close due to the fact that the edge cloud computing is dominant for energy saving in this case.

VI. Conclusion

In this paper, we studied the joint design of computing services when edge cloud computing and central cloud computing coexist in a two-tier HetNet with MIMO or massive MIMO self-backhaul. By jointly optimizing the cloud selection, the UEs' transmit powers, the SBSs' receive beamforming vectors and the transmit covariance matrices, the network's total energy consumption for task offloading and computation can be minimized while meeting both the edge processing and central processing (backhaul) latency constraints. An iterative algorithm was proposed to solve the formulated non-convex mixed-integer optimization problem, which can ensure the convergence and that better performance can be achieved than any existing feasible solutions. The simulation results have further confirmed that the proposed solution can greatly enhance the system performance, especially compared with the edge-cloud-only and central-cloud-only computing schemes, indicating the great value of cooperation between edge and central clouds. Moreover, we showed that the massive MIMO backhaul can largely decrease the complexity of the proposed algorithm while achieving even better performance.

Appendix A
Proof of Lemma 1

Based on problem (15), we can easily find that each SBS's receive beamformer $\mathbf{w}_n$ aims to maximize the SINR, i.e.,

$$\max_{\mathbf{w}_n} \ \gamma_n^{(\mathrm{a})}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n). \tag{A.1}$$

Problem (A.1) can be rewritten as

$$\max_{\mathbf{w}_n} \ \frac{\mathbf{w}_n^H \mathbf{\Omega}_n \mathbf{w}_n}{\mathbf{w}_n^H \mathbf{\Omega}_{-n} \mathbf{w}_n}. \tag{A.2}$$

Note that (A.2) is a generalized eigenvector problem and the optimal $\mathbf{w}_n^*$ is the corresponding eigenvector associated with the largest eigenvalue of the matrix $(\mathbf{\Omega}_{-n})^{-1}\mathbf{\Omega}_n$. Thus, we obtain the result in (16).

Appendix B
Proof of Lemma 2

The Lagrange function of problem (17) is

$$\begin{aligned}\mathcal{L}(\mathbf{p}^{\mathrm{u}}, \mathbf{t}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}) = &\sum_{n=1}^{N} I_n t_n + \sum_{n=1}^{N} \lambda_n \left(p_n^{\mathrm{u}} - t_n R_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n)\right) \\ &+ \sum_{n=1}^{N} \mu_n \left(\tau_n - \gamma_n^{\mathrm{a}}(\mathbf{p}^{\mathrm{u}}, \mathbf{w}_n)\right) \\ &+ \sum_{n=1}^{N} \nu_n \left(p_n^{\mathrm{u}} - P_{\max}^{\mathrm{u}}\right), \end{aligned} \tag{B.1}$$

where $\{\lambda_n, \mu_n, \nu_n\}_{n=1}^N$ are non-negative Lagrange multipliers. Based on the definition of KKT conditions, we have

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial p_n^{\mathrm{u}}} = &\lambda_n - \lambda_n t_n \frac{\partial R_n^{\mathrm{a}}}{\partial p_n^{\mathrm{u}}} - \mu_n \frac{\partial \gamma_n^{\mathrm{a}}}{\partial p_n^{\mathrm{u}}} + \nu_n \\ &- \sum_{j \neq n}^{N} \lambda_j t_j \frac{\partial R_j^{\mathrm{a}}}{\partial p_n^{\mathrm{u}}} - \sum_{j \neq n}^{N} \mu_j \frac{\partial \gamma_j^{\mathrm{a}}}{\partial p_n^{\mathrm{u}}} = 0, \end{aligned} \tag{B.2}$$

$$\frac{\partial \mathcal{L}}{\partial t_n} = I_n - \lambda_n R_n^{\mathrm{a}} = 0, \tag{B.3}$$

$$\lambda_n \left(p_n^{\mathrm{u}} - t_n R_n^{\mathrm{a}}\right) = 0, \tag{B.4}$$

$$\mu_n \left(\tau_n - \gamma_n^{\mathrm{a}}\right) = 0, \tag{B.5}$$

$$\nu_n \left(p_n^{\mathrm{u}} - P_{\max}^{\mathrm{u}}\right) = 0. \tag{B.6}$$

In (B.2), we have $\frac{\partial R_j^{\mathrm{a}}}{\partial p_n^{\mathrm{u}}} = -\frac{B^{\mathrm{a}}}{\ln 2} \frac{\left(\gamma_j^{\mathrm{a}}\right)^2 |\mathbf{w}_j^H \mathbf{h}_{n,j}^{\mathrm{a}}|^2}{p_j^{\mathrm{u}} |\mathbf{w}_j^H \mathbf{h}_{j,j}^{\mathrm{a}}|^2 \left(1+\gamma_j^{\mathrm{a}}\right)}$, and $\frac{\partial \gamma_j^{\mathrm{a}}}{\partial p_n^{\mathrm{u}}} = -\frac{\left(\gamma_j^{\mathrm{a}}\right)^2 |\mathbf{w}_j^H \mathbf{h}_{n,j}^{\mathrm{a}}|^2}{p_j^{\mathrm{u}} |\mathbf{w}_j^H \mathbf{h}_{j,j}^{\mathrm{a}}|^2}$. Based on (B.2)–(B.6), we observe that the N subproblems shown in (18) has the same KKT conditions with problem (17). In other words, problems (17) and (18) have the same optimal solution. In addition, since $R_n^{\mathrm{a}} > 0$, we have $\lambda_n = \frac{I_n}{R_n^{\mathrm{a}}} > 0$ based on (B.3), and then $t_n = \frac{p_n^{\mathrm{u}}}{R_n^{\mathrm{a}}}$ based on (B.4). Likewise, by considering the KKT conditions of N subproblems in (18), we find that they are identical to those shown in (B.2)–(B.6).

Appendix C
Proof of Theorem 1

Based on (B.2), (B.5) and (B.6) of Appendix B, KKT conditions for subproblem (18) is given by

$$\lambda_n + M_n - \frac{B_{\mathrm{a}}}{\ln 2} \frac{\lambda_n t_n \Lambda_n}{1 + \gamma_n^{\mathrm{a}}} - \mu_n \Lambda_n + \nu_n = 0, \tag{C.1}$$

$$\mu_n \left(\tau_n - \gamma_n^{\mathrm{a}}\right) = 0, \tag{C.2}$$

$$\nu_n \left(p_n^{\mathrm{u}} - P_{\max}^{\mathrm{u}}\right) = 0, \tag{C.3}$$

where $\Lambda_n = \frac{|\mathbf{w}_n^H \mathbf{h}_{n,n}^{\mathrm{a}}|^2}{\sum_{i=1, i\neq n}^{N} p_i^{\mathrm{u}} |\mathbf{w}_n^H \mathbf{h}_{i,n}^{\mathrm{a}}|^2 + |\mathbf{w}_n^H \mathbf{n}_i|^2}$. From (C.1) and the definition of $\gamma_n^{\mathrm{a}} = p_n^{\mathrm{u}} \Lambda_n$ in (2), we see that the optimal $p_n^{\mathrm{u}*}$ meets

$$p_n^{\mathrm{u}*} = \frac{B_{\mathrm{a}}}{\ln 2} \frac{\lambda_n t_n}{\lambda_n + M_n - \mu_n^* \Lambda_n + \nu_n^*} - \frac{1}{\Lambda_n}, \tag{C.4}$$

where μ_n^* and ν_n^* satisfy the KKT conditions (C.2) and (C.3), respectively. To explicitly obtain $\{p_n^{\mathrm{u}*}, \mu_n^*, \nu_n^*\}$, we need to consider the following cases:

- Case 1: When $p_n^{\mathrm{u}*} \in \left(\frac{\tau_n}{\Lambda_n}, P_{\max}^{\mathrm{u}}\right)$, $\mu_n^* = \nu_n^* = 0$ according to (C.2) and (C.3). In this case, $p_n^{\mathrm{u}*} = G_n$ with $G_n = \frac{B_{\mathrm{a}}}{\ln 2} \frac{\lambda_n t_n}{\lambda_n + M_n} - \frac{1}{\Lambda_n}$ according to (C.4). Therefore, if $G_n \in \left[\frac{\tau_n}{\Lambda_n}, P_{\max}^{\mathrm{u}}\right]$, $p_n^{\mathrm{u}*} = G_n$ and $\mu_n^* = \nu_n^* = 0$.
- Case 2: If $G_n < \frac{\tau_n}{\Lambda_n}$, it is seen from (C.4) that $\mu_n^* > 0$. In this case, $p_n^{\mathrm{u}*} = \frac{\tau_n}{\Lambda_n}$ and $\nu_n^* = 0$ according to (C.2) and (C.3). Substituting $p_n^{\mathrm{u}*} = \frac{\tau_n}{\Lambda_n}$ and $\nu_n^* = 0$ into (C.4), we obtain $\mu_n^* = \frac{\lambda_n + M_n}{\Lambda_n} - \frac{B_{\mathrm{a}}}{\ln 2} \frac{\lambda_n t_n}{\tau_n + 1}$
- Case 3: If $G_n > P_{\max}^{\mathrm{u}}$, it is seen from (C.4) that $\nu_n^* > 0$. In this case, $p_n^{\mathrm{u}*} = P_{\max}^{\mathrm{u}}$ and $\mu_n^* = 0$ according to (C.3) and (C.2). Substituting $p_n^{\mathrm{u}*} = P_{\max}^{\mathrm{u}}$ and $\mu_n^* = 0$ into (C.4), we obtain $\nu_n^* = \frac{B_{\mathrm{a}}}{\ln 2} \frac{\lambda_n t_n}{P_{\max}^{\mathrm{u}} + 1/\Lambda_n} - \lambda_n - M_n$.

Thus, we get the optimal $\{p_n^{\mathrm{u}*}, \mu_n^*, \nu_n^*\}$ as shown in **Theorem** 1.

References

[1] A. U. R. Khan, M. Othman, S. A. Madani, and S. U. Khan, "A survey of mobile cloud computing application models," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 393–413, 1st Quart., 2013.

[2] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Commun. Mobile Comput.*, vol. 13, no. 18, pp. 1587–1611, Dec. 2013.

[3] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45–55, Nov. 2014.

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

[5] A. Reznik *et al.*, "Cloud RAN and MEC: A perfect pairing," ETSI, Sophia Antipolis, France, White Paper 23, Feb. 2018.

[6] D. Sabella *et al.*, "Toward fully connected vehicles: Edge computing for advanced automotive communications," 5G Automot. Assoc. (5GAA), White Paper, Dec. 2017.

[7] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[8] M. Chiang, S. Ha, C.-L. I, F. Risso, and T. Zhang, "Clarifying fog computing and networking: 10 questions and answers," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 18–20, Apr. 2017.

[9] K. Dolui and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," in *Proc. Global Internet Things Summit (GIoTS)*, Jun. 2017, pp. 1–6.

[10] T. Verbelen, P. Simoens, F. De Turck, and B. Dhoedt, "Cloudlets: Bringing the cloud to the mobile user," in *Proc. ACM Workshop Mobile Cloud Comput. Services*, 2012, pp. 29–36.

[11] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.

[12] X. Lyu *et al.*, "Optimal schedule of mobile edge computing for Internet of Things using partial information," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2606–2615, Nov. 2017.

[13] E. Cuervo *et al.*, "MAUI: Making smartphones last longer with code offload," in *Proc. ACM MobiSys*, San Francisco, CA, USA, Mar. 2010, pp. 49–62.

[14] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 945–953.

[15] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 6, pp. 1107–1117, Jun. 2013.

[16] S. Deng, L. Huang, J. Taheri, and A. Y. Zomaya, "Computation offloading for service workflow in mobile cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, pp. 3317–3329, Dec. 2015.

[17] O. Munoz, A. Pascual-Iserte, and J. Vidal, "Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4738–4755, Oct. 2015.

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

[19] 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.

[20] A. Al-Shuwaili and O. Simeone, "Energy-efficient resource allocation for mobile edge computing-based augmented reality applications," *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 398–401, Jun. 2017.

[21] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.

[22] X. Sun and N. Ansari, "Green cloudlet network: A sustainable platform for mobile cloud computing," *IEEE Trans. Cloud Comput.*, to be published.

[23] 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.

[24] X. Hu, K.-K. Wong, and K. Yang, "Wireless powered cooperation-assisted mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2375–2388, Apr. 2018.

[25] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Mar. 2018.

[26] S. Bi and Y. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, Jun. 2018.

[27] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, "Challenges on wireless heterogeneous networks for mobile cloud computing," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 34–44, Jun. 2013.

[28] T. Zhao, S. Zhou, X. Guo, Y. Zhao, and Z. Niu, "A cooperative scheduling scheme of local cloud and Internet cloud for delay-aware mobile cloud computing," in *Proc. IEEE Globecom Commun. Conf. Workshops (GC WKSHPS*), San Diego, CA, USA, Dec. 2015, pp. 1–6.

[29] F. Ben Jemaa, G. Pujolle, and M. Pariente, "QoS-aware VNF placement optimization in edge-central carrier cloud architecture," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC USA, Dec. 2016, pp. 1–7.

[30] M.-H. Chen, M. Dong, and B. Liang, "Joint offloading decision and resource allocation for mobile cloud with computing access point," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 3516–3520.

[31] D. Pompili, A. Hajisami, and H. Viswanathan, "Dynamic provisioning and allocation in cloud radio access networks (C-RANs)," *Ad Hoc Netw.*, vol. 30, pp. 128–143, Jul. 2015.

[32] D. Liu *et al.*, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart. 2016.

[33] 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, USA, Jun. 2010, p. 19.

[34] S. Melendez and M. P. McGarry, "Computation offloading decisions for reducing completion time," in *Proc. IEEE CCNC*, Las Vegas, NV, USA, Jan. 2017, pp. 160–164.

[35] L. Yang, J. Cao, Y. Yuan, T. Li, A. Han, and A. Chan, "A framework for partitioning and execution of data stream applications in mobile cloud computing," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 4, pp. 23–32, Mar. 2013.

[36] T. Han *et al.*, "Small cell offloading through cooperative communication in software-defined heterogeneous networks," *IEEE Sensors J.*, vol. 16, no. 20, pp. 7381–7392, Oct. 2016.

[37] D. Valocchi, D. Tuncer, M. Charalambides, M. Femminella, G. Reali, and G. Pavlou, "SigMA: Signaling framework for decentralized network management applications," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 3, pp. 616–630, Sep. 2017.

[38] Y. Yang and M. Pesavento, "A unified successive pseudoconvex approximation framework," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3313–3328, Jul. 2017.

[39] A. Zappone and E. Jorswieck, *Energy Efficiency in Wireless Networks Via Fractional Programming Theory*, vol. 11, nos. 3–4. Boston, MA, USA: Now, 2015.

[40] M. Grant, S. Boyd, and Y. Ye. (2008). *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available: https://web.stanford.edu/~boyd/papers/pdf/disc_cvx_prog.pdf

[41] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[42] H. Q. Ngo and E. G. Larsson, "No downlink pilots are needed in TDD massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2921–2935, May 2017.

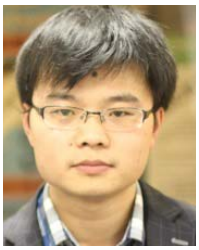
[43] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[44] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[45] A. He, L. Wang, Y. Chen, K.-K. Wong, and M. Elkashlan, "Spectral and energy efficiency of uplink D2D underlaid massive MIMO cellular networks," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3780–3793, Sep. 2017.

[46] *Technical Specification Group Radio Access Network: Evolved Universal Terrestrial Radio Access (E-UTRA): Further Advancements for E-UTRA Physical Layer Aspects*, document TR 36.814. 3GPP, Mar. 2017.

Xiaoyan Hu (S'16) received the M.Eng. degree in information and communication engineering from Xi'an Jiaotong University, China, in 2016. She is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering, University College London, U.K. Her research interests are in the areas of mobile edge computing, UAV communications, wireless energy harvesting, cooperative communications, and physical-layer security. She was selected as an Exemplary Reviewer of the IEEE COMMUNICATIONS LETTERS in 2017.

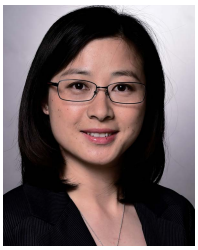


Lifeng Wang (M'16) received the Ph.D. degree in electronic engineering from the Queen Mary University of London. He was the Research Associate with the Department of Electronic and Electrical Engineering, University College London (UCL). He joined the Department of Electrical Engineering, Fudan University. His research interests include massive MIMO, millimeter wave, dense HetNets, edge caching, physical-layer security, and wireless energy harvesting. He received the Exemplary Editor Certificate of the IEEE COMMUNICATIONS LETTERS from 2016 to 2018.

Kai-Kit Wong (M'01–SM'08–F'16) received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively.

After graduation, he took up academic and research positions at The University of Hong Kong, Lucent Technologies, Bell-Labs, Holmdel, the Smart Antennas Research Group, Stanford University, and the University of Hull, U.K. He is currently the Chair in Wireless Communications with the Department of Electronic and Electrical Engineering, University College London, U.K. His current research centers around 5G and beyond mobile communications, including topics such as massive MIMO, full-duplex communications, millimeter-wave communications, edge caching and fog networking, physical-layer security, wireless power transfer and mobile computing, V2X communications, and, of course, cognitive radios. There are also a few other unconventional research topics that he has set his heart on, including, for example, fluid antenna communications systems and remote ECG detection.

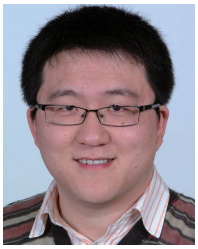
Dr. Wong is also a fellow of IET and is also on the editorial board of several international journals. He was a co-recipient of the 2013 IEEE Signal Processing Letters Best Paper Award, the 2000 IEEE VTS Japan Chapter Award at the IEEE Vehicular Technology Conference in Japan in 2000, and a few other international best paper awards. He has been serving as a Senior Editor for the IEEE COMMUNICATIONS LETTERS since 2012 and the IEEE WIRELESS COMMUNICATIONS LETTERS since 2016. He has served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 2009 to 2012 and an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2005 to 2011. He was also a Guest Editor of the IEEE JSAC SI on virtual MIMO in 2013. He is also a Guest Editor of the IEEE JSAC SI on physical-layer security for 5G.



Meixia Tao (S'00–M'04–SM'10–F'19) received the B.S. degree in electronic engineering from Fudan University, Shanghai, China, in 1999, and the Ph.D. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology in 2003.

She was a member of Professional Staff with the Hong Kong Applied Science and Technology Research Institute from 2003 to 2004 and a Teaching Fellow and then an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, from 2004 to 2007. She is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China. Her current research interests include wireless caching, edge computing, physical-layer multicasting, and resource allocation.

Dr. Tao was a recipient of the IEEE Marconi Prize Paper Award in 2019, the IEEE Heinrich Hertz Award for Best Communications Letters in 2013, the IEEE/CIC International Conference on Communications in China (ICCC) Best Paper Award in 2015, and the International Conference on Wireless Communications and Signal Processing (WCSP) Best Paper Award in 2012. She received the IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award in 2009. She has served on the Editorial Board of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS as a member of the Executive Editorial Committee from 2015 to 2019 and as an Editor from 2007 to 2011. She has served as the Symposium Oversight Chair of IEEE ICC 2019, the Symposium Co-Chair of IEEE GLOBECOM 2018, the TPC Chair of IEEE/CIC ICCC 2014, and the Symposium Co-Chair of IEEE ICC 2015. She was on the Editorial Board of the IEEE TRANSACTIONS ON COMMUNICATIONS from 2012 to 2018, the IEEE COMMUNICATIONS LETTERS from 2009 to 2012, and the IEEE WIRELESS COMMUNICATIONS LETTERS from 2011 to 2015.



Yangyang Zhang received the B.S. and M.S. degrees in electronics and information engineering from Northeastern University, Shenyang, China, in 2002 and 2004, respectively, and the Ph.D. degree in electrical engineering from the University of Oxford, Oxford, U.K., in 2008.

He is currently the Executive Vice President of the Kuang-Chi Institute of Advanced Technology, China. His research interests include multiple-input–multiple-output wireless communications and stochastic optimization algorithms. He has been awarded more than 20 honors. He also authored or coauthored more than 30 refereed articles.

Zhongbin Zheng received the bachelor's and master's degrees in information and communications engineering from the Beijing University of Posts and Telecommunications in 2002 and 2005, respectively. He was the former Head of the Technology Department, East China Institute, Ministry of Industry and Information Technology. He is currently the Vice Director of the China Academy of Information and Communications Technology and the East China Institute of Telecommunications. He is very active in research, resulting in not only a number of international paper publications but also patents and draft standards.