Task-Oriented Delay-Aware Multi-Tier Computing in Cell-free Massive MIMO Systems

Kunlun Wang, Dusit Niyato, Wen Chen, and Arumugam Nallanathan

***Abstract*—Multi-tier computing can enhance the task computation by multi-tier computing nodes. In this paper, we propose a cell-free massive multiple-input multiple-output (MIMO) aided computing system by deploying multi-tier computing nodes to improve the computation performance. At first, we investigate the computational latency and the total energy consumption for task computation, regarded as total cost. Then, we formulate a total cost minimization problem to design the bandwidth allocation and task allocation, while considering realistic heterogenous delay requirements of the computational tasks. Due to the binary task allocation variable, the formulated optimization problem is non-convex. Therefore, we solve the bandwidth allocation and task allocation problem by decoupling the original optimization problem into bandwidth allocation and task allocation subproblems. As the bandwidth allocation problem is a convex optimization problem, we first determine the bandwidth allocation for given task allocation strategy, followed by conceiving the traditional convex optimization strategy to obtain the bandwidth allocation solution. Based on the asymptotic property of received signal-to-interference-plus-noise ratio (SINR) under the cell-free massive MIMO setting and bandwidth allocation solution, we formulate a dual problem to solve the task allocation subproblem by relaxing the binary constraint with Lagrange partial relaxation for heterogenous task delay requirements. At last, simulation results are provided to demonstrate that our proposed task offloading scheme performs better than the benchmark schemes, where the minimum-cost optimal offloading strategy for heterogeneous delay requirements of the computational tasks may be controlled by the asymptotic property of the received SINR in our proposed cell-free massive MIMO-aided multi-tier computing systems.**

***Index Terms*—Multi-tier computing systems, cell-free massive MIMO systems, energy and delay tradeoff, delay requirements, task offloading**

I. Introduction

SINCE there are exponential growth of mobile devices in the networks, wireless traffic is growing tremendously recently. It is expected that and the increasing amount of traffic will continue to grow steadily in the coming years. With the proliferation of wireless traffic, delay and energy consumption have emerged as key design metrics for wireless communication systems [1], [2]. Additionally, more and more smart devices are connected to the wireless network with the development of intelligent Internet of Things (IoT), it is estimated that in excess of 24.6 billion connected devices by 2025 [3]. Meanwhile, the significant growth of novel intelligent applications with intensive tasks (e.g., AR/VR) typically require ultra-reliable and low-latency communications (URLLC) and demand efficient power management for realtime task processing and high energy efficiency (EE) [4], [5]. However, mobile hand-held devices have limited computation, energy and storage resources, as well as limited battery capacity due to their compact form-factor. These defects pose critical challenges for the realtime intelligent applications. By enabling flexible computation, storage and communication resource coordination, multi-tier computing is a novel and efficient computing architecture, which can schedule intensive tasks to multi-tier computing servers at heterogeneous base stations (BSs) in the edge/fog or cloud of wireless communication systems [6], [7].

From the perspective of EE, the energy consumption model adopted in most of related works is over-simplified, which only models the transmission energy or task computational energy [7], [8]. However, a more general energy model is not negligible in a multi-tier computing system considering various computation and communication energy consumptions at the multi-tier nodes. From the perspective of task computation delay, the computation delay model proposed in most of related works fails to capture the effects of heterogeneous delay requirements. There are diversified applications in multi-tier computing systems, some tasks are delay tolerant, while some tasks are delay sensitive. However, most of these works fail to consider joint influence of the heterogeneous delay requirements and the total energy consumption, which can be also regarded one of the key task computation metrics in next generation wireless networks.

Regarded as one of the key technologies for next generation wireless communication systems, massive multiple-input multiple-output (MIMO) is capable of significantly improving the task offloading rate so as to improve the task compuation efficiency [9]–[11]. In order to make full use of the benefits of massive MIMO in a multi-tier computing system, the cell-free massive MIMO based multi-tier computing is considered. Cell-free massive MIMO is the network-centric massive MIMO, which is distributed across the network [12], [13]. Thus, a cell-free massive MIMO system supports multiple number of antennas distributed over a large number of access points (APs) in a network, where each AP can serve a small number of devices. These APs are coordinated by the central processing unit (CPU) through a high-transmission data rate

K. Wang is with the Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China, and also with the School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China (e-mail: klwang@cee.ecnu.edu.cn).

D. Niyato is with School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: dniyato@ntu.edu.sg).

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

A. Nallanathan is with the School of Electronic Engineering and Computer Science at Queen Mary University of London, UK. (email: a.nallanathan@qmul.ac.uk).

and ultra-reliable backhaul link. For implementing multi-tier computing, we assume that each AP and the CPU are respectively equipped with an independent computing server, while the computing server for CPU has much larger computation capacity compared to that of AP. Each AP serves all devices in its coverage area. Thus, each device is capable of utilizing the abundant computational resources on the CPU or one of its connected APs (i.e., multi-tier computing framework). In this context, our proposed computing framework focuses on multi-tier computing nodes (CNs) in cell-free massive MIMO systems, e.g., fog access nodes (FANs), cloud access nodes (CANs) and CPU, realizing collaborative task computing. These CNs are able to help to execute the computational tasks offloaded from task nodes (TNs) according to their computing capabilities, task delay requirements and total cost.

A. Related Work

In order to address the inefficient task computation issues, multi-tier computing is capable of offloading the intensive tasks to multi-tier nearby CNs with the powerful computing capability realizing remote task execution [14]. Therefore, task offloading has received more and more research attention in different edge/fog computing scenarios [6], [7], [9]–[11], [15]–[17]. In particular, Wang *et al.* [9] proposed a massive MIMO-aided task offloading system, where multiple TNs can schedule their tasks to nearby computing nodes (CNs) by a massive MIMO-based FAN. As an extension of this work, Wang *et al.* [10] proposed a relay assisted multi-tier computing system equipped with massive antennas to improve the task computation performance, they investigated the design of the task allocation, service caching and power allocation jointly to minimize the total task computation latency. Additionally, Wang *et al.* [6] investigated a non-orthogonal multiple access (NOMA)-assisted task offloading system for industrial Internet of things (IIoT), where TNs offload their intensive tasks to nearby CNs via NOMA technique for task computation. Also Wang *et al.* [11] proposed a task offloading framework in a intelligent reflecting surface (IRS) and massive MIMO relay assisted multi-tier computing system, where multiple TNs can offload their intensive tasks to nearby massive MIMO relay node (MRN) and FAN via the IRS technique for task execution in CNs. Liu *et al.* [18], [19] proposed a mapping framework for task offloading by mapping multiple tasks or TNs into multiple HNs, where they studied a generalized Nash equilibrium problem to minimize task's offloading delay in a distributed manner.

By equipping with a very large number of distributed APs in a network, cell-free massive MIMO is capable of significantly improving network-throughput as well as EE [20]. Given the benefits of cell-free massive MIMO, the integration of multi-tier computing and cell-free massive MIMO can improve the task offloading performance in a multi-tier computing system [20]–[22]. There are some works on resource management of task offloading in a cell-free massive MIMO system in recent years. Mukherjee *et al.* [21] proposed an edge computing-assisted cell-free massive MIMO architecture, where the edge servers and cloud are located at each AP and the central server of this system, respectively, and the authors analysis the task offloading performance by devising suitable communication resource allocation and task allocation strategies. Ke *et al.* [22] introduced a grant-free massive access IoT system, where multiple cooperative APs serve massive devices in the network via cell-free massive MIMO technique, they studied two computation strategies at the CNs for massive devices access, i.e., cloud task computation and edge task computation. Wang *et al.* [10] investigated the joint strategy of task offloading, computational task caching and power allocation in edge computing systems to minimize the total task offloading latency.

Although the above works have revealed the benefits of cell-free massive MIMO-assisted edge computing [9]–[11], the energy-delay tradeoff in resource management and multi-tier task computation have not been considered. By considering task computation energy and latency costs in a multi-tier task offloading framework, scheduling tasks to multi-tier CNs can reduce the congestion of task computation as well as reducing the computation energy consumption of each user. To exploit the benefits of energy-delay tradeoff in edge computing framework, there are some works being invested into online dynamic tasks allocation with energy harvesting [23], task offloading of mobile devices formulated as a constrained multi-objective optimization problem on minimizing both the task computation energy consumption and task computation latency [24], and jointly task offloading and resource management optimization [25], as well as minimizing task response time and packet losses to improve the realtime performance and reliability of task processing [26]. However, the influence of heterogeneous delay requirements of computational tasks has not been studied, which represents the different delay requirements of diverse novel applications. Additionally, the influence of heterogeneous delay requirements of the tasks for task allocation in multi-tier computing nodes has not been studied either. Furthermore, all the existing works consider the uncoordinated distributed edge computing scenario. Thanks to the rapid development of cell-free massive MIMO [27], the task offloading via cell-free massive MIMO will be increasingly adopted in multi-tier computing framework.

B. Main Contributions

Although the above contributions have revealed the benefits of task offloading in a cell-free massive MIMO framework, the multi-tier collaborative task computation minimizing the total energy consumption and latency in cell-free massive MIMO frameworks has not been considered to the best of our knowledge. The influence of delay and energy weight has not been well studied either, which can be regarded as the heterogeneous delay requirements of the tasks and will be a key performance metric for multi-tier computing systems [6], [30]–[32], and the heterogeneous delay requirements is particularly important for battery-limited mobile devices running diverse novel applications. The proposed cell-free massive MIMO-aided multi-tier computing framework includes heterogeneous CNs, e.g., mobile devices, fog/cloud access points, and cloud. The total energy consumption consist of both the task transmission energy and task computation energy, and the total

TABLE I
NOVELTY COMPARISON

	[8]-2016	[15]-2019	[28]-2017	[29]-2019	[6]-2020	[9]-2020	[10]-2022	[11]-2022	Our work
Joint task allocation and communication resource allocation	✓	✓	✓		✓	✓	✓	✓	✓
Massive MIMO				✓		✓	✓	✓	✓
Minimizing task computation time	✓								✓
Multi-tier computing					✓	✓	✓	✓	✓
Cell-free massive MIMO									✓
Energy-delay tradeoff									✓
Heterogeneous delay requirements									✓

latency cost includes both the task transmission delay and task computation delay. The task computation energy includes the energy consumed by both the local devices and the remote CN. Furthermore, to characterize the effects of different delay requirements for the tasks in multi-tier computing systems, we employ the delay and energy weight for determining the task allocation policy. That is, delay sensitive and tolerant tasks are classified into different categories by the weights. Based on the requirements of delay sensitive and tolerant tasks, we aim for studying how cell-free massive MIMO systems tackle the challenge of task allocation. In the expressions for the task computation latency and energy consumption, we obtain that the bandwidth allocation variable and task offloading variable are separated. Then, the task allocation and the bandwidth allocation constraints are separable. Therefore, we decouple the original task offloading Problem into two sub-problems: bandwidth allocation optimization and task allocation optimization. Specifically, we first exploit the optimal bandwidth allocation strategy to minimize the total cost of our proposed cell-free massive MIMO-assisted multi-tier computing systems. Secondly, we characterize the relationship between cell-free massive MIMO and task allocation strategy based on the asymptotic property of the signal-to-interference-plus-noise ratio (SINR) with very large number of APs. Finally, we optimize the task allocation for heterogeneous delay requirements of the tasks. We have explicitly compared our unique contributions with the state-of-the-art, which are shown in Table I and further summarized as follows:

- In terms of multi-tier task offloading system model, a novel cell-free massive MIMO assisted multi-tier computing framework is proposed for task offloading, which consists of a group of a large number of distributed APs with sharable computing resources to execute the tasks. Furthermore, a problem minimizing the total energy consumption and latency cost is formulation, which offers a new approach to task offloading in a multi-tier computing system.
- In terms of task offloading optimization and resource management analysis, an adaptive bandwidth allocation strategy is proposed for minimizing the total energy and delay cost, where the most efficient bandwidth for each task offloading link can be determined according to the dynamic channels. As far as we know, considering the bandwidth allocation in a cell-free massive MIMO-assisted task offloading system has not been studied recently.
- Furthermore, we optimize the task allocation strategy based on the bandwidth allocation result. Given the delay sensitive and tolerant tasks, we study how cell-free massive MIMO systems tackle the challenge of task allocation with heterogeneous delay requirements. Then, we employ the delay and energy weight to determine the optimal task allocation policy.
- Finally, the performances of the proposed bandwidth allocation and task offloading strategy have been evaluated through simulations relying on diverse system parameters. The simulation results demonstrate that our task offloading strategy achieves significantly performance improvement in total cost compared to the benchmark schemes subject to realistic communication and computation constraints.

C. Paper Organization

Our paper is organized as follows. The system model is introduced in Section II, while the problem formulation and analysis are presented in Section III. In Section IV, we optimize the bandwidth and task allocations in terms of heterogeneous delay requirements by the asymptotic property of the massive MIMO to minimize the total energy and delay cost in cell-free massive MIMO-assisted multi-tier computing systems. our simulation results are shown in Section V. Finally, conclusions are provided in Section VI. Table II lists the notations.

II. SYSTEM MODEL

A. Multi-tier Computing Network

A cell-free massive MIMO-aided multi-tier computing network is shown in Fig. 1, which consists of K TNs with computational tasks (e.g., artificial intelligence model training), K FANs and M CANs, where the active TNs rely on node to node (N2N) communications for task offloading to FANs. The FANs are connected to the CNs by backhaul links using high-rate ultra-reliable optical fiber transmission. The TNs are communicated with the FAN by wireless links, and the FANs offload the tasks to the CANs, while the CANs are connected to the CPU by backhaul links using high-rate optical fiber. As the next generation wireless network is expected to satisfy the extensive quality of service (QoS) requirements

TABLE II
NOTATIONS

Definition	Notation	Definition	Notation
Bandwidth	B	Task size of task k	l_k
Bandwidth allocation variable for TN k	η_k	Task offloading decision variable of task node k	α_k
Transmit power of TN k	p_b	Variance of AWGN	σ_k^2
Transmit symbol of task node k	s_k	Received SINR of symbol x_k	γ_k
Task offloading time of first hop	t_k^{TN}	Offloading energy of first hop	E_k^{TN}
Task offloading time of second hop	t_k	Offloading energy of first hop	E_k^{FAN}
Computational latency at CN	$t_{k,\mathrm{comp}}^F$	Computational latency at CPU	$t_{k,\mathrm{comp}}^C$
Computational energy consumption at CN	$E_{\mathrm{re,k}}$	Total energy consumption of TN k	$E_{\mathrm{total,k}}$
Total task offloading latency of task node k	$T_{\mathrm{total,k}}$	Transmitted symbol from the kth FAN	x_k
Weights of delay and energy consumption	μ_k	Weighted sum cost of task node k	Ω_k
Total cost	Ω_{total}	Number of CAN	M

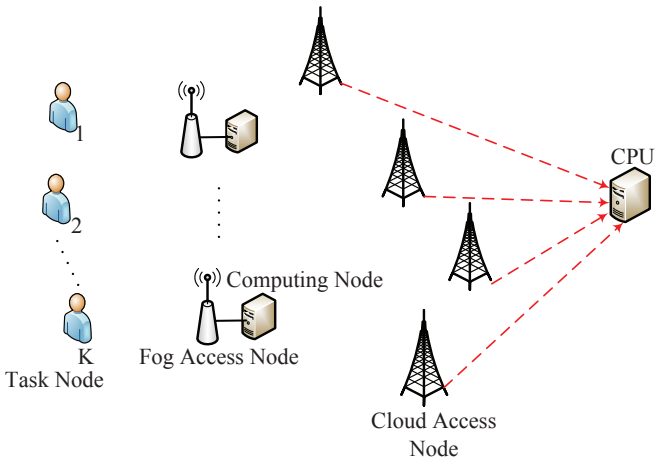

Fig. 1. Illustration of a cell-free massive MIMO-assisted multi-tier computing system consisting of K TNs, K fog access nodes (FANs) and M cloud access nodes (CANs), where TNs offload the tasks to the FANs or CANs for multi-tier computing.

for wireless communications, a cell-free massive MIMO-based multi-tier computing architecture is proposed. Multi-tier computing organizes and manages the task computation and transmission from multiple heterogeneous CNs. Under this circumstance, the tasks can be offloaded to the FAN and CPU with the help of CANs, which can improve the performance of the overall computation efficiency.

B. Transmission Model

We consider independent and identically distributed (i.i.d.) quasi-static Rayleigh fading. In particular, each inter-node channel remains invariant within one time slot, but varies independently across different time slots and links. To make full use of the spectrum, we consider that each TN occupies a part of the bandwidth to facilitate task offloading by frequency division multiple access (FDMA), which can avoid the TNs co-channel interference. Let B and $\eta_k \in [0,1]$ be the total available bandwidth for the links between TNs and FANs and the bandwidth allocation variable for TN k, respectively, recall that the bandwidth allocated to the ith TN is $\eta_k B$. We denote $\boldsymbol{\eta}$ as the bandwidth allocation vector, which can be shown as $\boldsymbol{\eta} = [\eta_1, \eta_2, \cdots, \eta_K]$.

We denote $\alpha_k \in \{0,1\}$ and $\mathbf{a} = [\alpha_1, \cdots, \alpha_K]$ as the offloading variable of TN k and the offloading vector, respectively. As we consider binary task offloading, we have $\alpha_k = 0$ if TN k executes its task locally by FAN computing, and $\alpha_k = 1$ if TN k executes its task by remote CPU computing. Let p_b be the transmit power of each TN, which is selected prior to the subcarrier allocation. Denote z_k as the additive white Gaussian noise (AWGN) with variance σ_k^2 at the kth FAN. Hence, the received task signal at the kth FAN is shown as

$$y_k = \sqrt{p_b}\sqrt{d_k}h_k x_k + z_k, \forall k \in \mathcal{K}, \tag{1}$$

where d_k and $h_k \sim \mathcal{CN}(0,1)$ represent the channel path loss and small scale fading of the link between the kth TN and the kth FAN. Let x_k be the transmit symbol of TN k. In addition, x_k satisfies $\mathsf{E}[|x_k|^2] = 1$.

Based on (1), the signal-to-noise ratio (SNR) of symbol x_k observed by the kth FAN can be expressed as

$$\gamma_{\mathrm{FAN,k}} = \frac{p_b d_k |h_k|^2}{\sigma_k^2}. \tag{2}$$

Then, based on the bandwidth allocation and received SNR in (2), the achievable task offloading rate of the first hop from the kth TN to the kth FAN is given by

$$r_k = \eta_k B \log_2\left(1 + \frac{p_b d_k |h_k|^2}{\sigma_k^2}\right). \tag{3}$$

Then, the task offloading time of the first hop is given by

$$t_k^{\mathrm{TN}} = \frac{l_k}{\eta_k B \log_2\left(1 + \frac{p_b d_k |h_k|^2}{\sigma_k^2}\right)}. \tag{4}$$

The corresponding energy consumption is given by

$$E_k^{\mathrm{TN}} = \frac{p_b l_k}{\eta_k B \log_2\left(1 + \frac{p_b d_k |h_k|^2}{\sigma_k^2}\right)}. \tag{5}$$

For the second hop, FAN k offloads its task to the CAN. By the cell-free transmissions, CAN receives all the tasks from the FANs. Then, the task offloading time and task offloading energy consumption of the second hop are respectively given by

$$t_k = \frac{\alpha_k l_k}{B \log_2(1+\gamma_k)}, \tag{6}$$

$$E_k^{\mathrm{FAN}} = \frac{q_k \alpha_k l_k}{B \log_2(1+\gamma_k)}, \tag{7}$$

where γ_k represents the received SINR for the TN k at CANs.

C. Computational Model

1) Local Computing: Consider that each TN has a computational task for the requested service, and hence we denote k_s as the task of TN k, which can be specified by the task size with l_k bits. For simplicity of analysis, we assume that the task for each TN is generated instantly, and the task offloading latency is from the arrival in the TN. Regrading task computation in TN, we assume that each TN does not have enough computation capacity to execute the task, the task needs to be offloaded to execute on nearby FAN. If the task is still computed locally, the execution delay could be largely due to the limited computing resources at the TN [33].

In our proposed multi-tier computing system, the total number of computing cycles of processing core is considered to be linearly proportional with the task size of each task to be processed [6], [9]. Let C_k represent the number of computing cycles required for executing 1-bit of input task at CN, and hence the total number of computing cycles required for executing the task from TN k is $C_k l_k$, which depends both on the type of processing core and on the task to be executed. Therefore, the task computational latency for k_s at CN is given by

$$t^F_{k,\text{comp}} = \frac{C_k(1-\alpha_k)l_k}{f^F_k}, \tag{8}$$

where f^F_k represents the core computing frequency at the CN.

2) Task Offloading: In terms of task offloading, where the task is offloaded to be processed by remote CPU. In this case, the TN transmits the task to the CPU through two hops wireless links, i.e, TNs to FANs link and FANs to CANs link. For the sake of simplicity, we assume that the CPU has multi-core, and each particular offloaded task can be independently assigned to a core Furthermore, each core is assumed to have the same maximum computing frequency $f^{\max}_0$ (in cycles per second). Therefore, we have computational latency of the task k_s at CPU, which is given by

$$t^C_{k,\text{comp}} = \frac{C_k\alpha_k l_k}{f^C_k}, \tag{9}$$

where f^C_k, $k \in \mathcal{K}$ represents the each core computing frequency in the CPU. Benefit from dynamic voltage and frequency scaling techniques (DVFS) [34], we assume that f^C_k is adjustable.

III. Problem Formulation and Analysis

A. Energy Consumption and Delay Analysis

Following the model in [8], let P_{CN} denote the computing energy consumption for each cycle at CN for local task computing. Then, $C_k P_{\text{CN}}$ represents the computing energy consumption per bit of each task. According to the task allocation decision, there are $(1-\alpha_k)l_k$-bits input task required to be processed at kth CN. Thus, the computational energy consumption of the task k_s at the k-th CN for local task computing is given by

$$E_{\text{re},k} = C_k(1-\alpha_k)l_k P_{\text{CN}}. \tag{10}$$

Specifically, the computational energy consumption of the task k_s at CPU is not considered, as the energy capacity of CPU is very large. Then, the total computational energy consumption for task k_s is given by

$$E_{\text{com},k} = E_{\text{re},k} = C_0(1-\alpha_k)l_k P_{\text{CN}}. \tag{11}$$

In all, the total task computation energy consumption for task k_s is composed of total task transmission energy and total task computation energy, which is given by

$$\begin{aligned} E_{\text{total},k} =& E^{\text{TN}}_k + E_{\text{re},k} + E^{\text{FAN}}_k \\ =& \frac{p_b l_k}{\eta_k B \log_2\left(1+\frac{p_{b,m} d_{b,m}|h_{b,m}|^2}{\sigma^2_m}\right)} + \\ & C_0(1-\alpha_k)l_k P_{\text{CN}} + \frac{q_k \alpha_k l_k}{B\log_2(1+\gamma_k)}. \end{aligned} \tag{12}$$

The total task offloading latency consists of the task computation delay plus the task transmission delay, and it is from the arrival in the TN. According to (4), (6), (8), and (9), the total task offloading latency of task k_s is given by

$$\begin{aligned} T_{\text{total,k}} =& t^{\text{TN}}_k + t^F_{k,\text{comp}} + t_k + t^C_{k,\text{comp}} \\ =& \frac{l_k}{\eta_k B \log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma^2_m}\right)} + \frac{C_k(1-\alpha_k)l_k}{f^F_k} \\ & + \frac{\alpha_k l_k}{B\log_2(1+\gamma_k)} + \frac{C_k\alpha_k l_k}{f^C_k}. \end{aligned} \tag{13}$$

B. Problem Formulation

In this section, the bandwidth and task allocations framework is discussed for heterogeneous delay requirements of the tasks under the cell-free massive MIMO setting, and the related optimization problem is formulated. Specifically, the total task computation cost in terms of latency cost and energy consumption minimization is formulated to improve the task execution efficiency of multi-tier task computation.

According to (12) and (13), the weighted sum cost for each TN k in terms of latency cost and energy consumption can be obtained, which is given by

$$\Omega_k = \mu_k E_{\text{total},k} + (1-\mu_k) T_{\text{total,k}}, \tag{14}$$

where μ_k and $(1-\mu_k)$ represent the weight of latency cost and energy consumption, respectively, and $0 \le \mu_k \le 1$. μ_k reflects the priority of different task requirements, under the condition of delay sensitive task, μ_k is small. On the other hand, if the task is delay tolerant, μ_k is large, which means total energy consumption is more important than delay cost for task computation. For simplicity of analysis, we set $\mu_k = 0$ for delay sensitive services, and we set $\mu_k = 1$ for delay tolerant services. Then, the total cost consists of computational energy consumption and computational delay, which can be formulated as

$$\begin{aligned} \Omega_{\text{total}} =& \sum_{k=1}^{K} w_k \Omega_k \\ =& \sum_{k=1}^{K} w_k \left\{\mu_k E_{\text{total},k} + (1-\mu_k) T_{\text{total,k}}\right\}, \end{aligned} \tag{15}$$

where w_k is the weighting factor for TN k.

Hence, the total cost minimization optimized problem of the proposed cell-free massive MIMO asisted multi-tier computing systems is expressed as

$$\min_{\boldsymbol{\eta},\boldsymbol{\alpha}} \quad \Omega_{\text{total}}, \tag{16a}$$

$$\text{s.t.} \quad 0 < \eta_k < 1, \forall k, \tag{16b}$$

$$\sum_{k=1}^{K} \eta_k = 1, \forall k, \tag{16c}$$

$$\alpha_k \in \{0, 1\}, \forall k. \tag{16d}$$

C. Problem Analysis

Using the above described problem formulation, we need to analyze Problem (16). Note that α_k is a binary variable, and hence Problem (16) has an non-convex feasible set. As binary variables has product-based relationship, the objective function of Problem (16) is also non-convex. As we know that it is challenging to find global optimum for a mixed discrete and non-convex optimization problem [35]. Since the feasible set and objective function are both non-convex, this problem is NP hard. Thus, we have to simplify Problem (16).

From (12) and (13), it can be observed that in the expressions for the task computation latency and energy consumption, the bandwidth allocation variable $\boldsymbol{\eta}$ and task offloading variable $\boldsymbol{\alpha}$ are separated. In this case, the task allocation and the bandwidth allocation constraints are separable. Therefore, we are capable of decoupling the original task offloading Problem (16) into two sub-problems: bandwidth allocation optimization (i.e., $\boldsymbol{\eta}$) and task allocation optimization (i.e., $\boldsymbol{\alpha}$). Note that the bandwidth allocation is based on the first hop task transmission from the TNs to the FANs, and the task allocation is based on the second hop task offloading from the FANs to CANs with the aid of cell-free massive MIMO.

IV. Cost Minimization Schemes

In this section, the bandwidth allocation and task offloading is optimized for heterogeneous delay requirements of the tasks by the asymptotic property of the massive MIMO to minimize the total task computation cost in our proposed cell-free massive MIMO-aided multi-tier computing systems.

A. Uplink Task Transmission

Consider that task transmissions from FANs to CANs are regarded as uplink task offloading, where all FANs send their received tasks to the CANs. We denote the channel fading coefficient between the FAN k and the CAN m as $g_{m,k}$, which can be expressed as [36]

$$g_{m,k} = \sqrt{\beta_{m,k}} h_{m,k}, \tag{17}$$

where $\beta_{m,k}$ denotes the large-scale fading and $h_{m,k} \sim \mathcal{CN}(0,1)$ denotes the small-scale channel fading between the link from the kth TN and the FAN m. Upon receiving the signal, all the FANs delivered their symbols simultaneously to the CAN, which can be expressed as

$$\mathbf{x} = \sqrt{\rho q_k}\mathbf{s}, \tag{18}$$

where $\mathbf{s} = [s_1, \cdots, s_K]^T$ denotes the transmission information-bearing signal vector with $\mathsf{E}(\mathbf{s}\mathbf{s}^\dagger) = \mathbf{I}_K$, and s_k ($\mathsf{E}[|s_k|^2] = 1$) denotes the signal transmitted from the kth TN to its paired FAN, and q_k denotes the task transmission power from the FAN k. Additionally, ρ represents the normalized uplink signal-to-noise ratio (SNR). The delivered symbol from FAN k can be expressed as

$$x_k = \sqrt{\rho q_k} s_k. \tag{19}$$

Note that each CAN receives aggregated signal from all the FANs, and the received symbol at the mth CAN can be expressed as

$$y_m = \sqrt{\rho}\sum_{k=1}^{K} g_{m,k}\sqrt{q_k}s_k + n_m, \mathbf{y}_m = \mathbf{g}_{m,k}\sqrt{\rho q_k}\mathbf{s} + \mathbf{n}_m, \tag{20}$$

where $n_m \sim \mathcal{CN}(0,1)$ denotes the noise at CAN m. Additionally, we employ a matched filtering strategy at the CANs. In this case, the received symbol can be weighted appropriately. To be more precisely, the received symbol y_m at the mth CAN need to be first multiplied by $\hat{\mathbf{g}}^*_{m,k}$. Let $\hat{\mathbf{g}}_m$ denote the estimated channel state information (CSI) of all FANs to mth CAN. Under this circumstance, the actual channel coefficient can be expressed as [37]

$$\mathbf{g}_m = \sqrt{1-\tau_D^2}\hat{\mathbf{g}}_m + \tau_D \boldsymbol{\Omega}_D, \tag{21}$$

where $\boldsymbol{\Omega}_D \in \mathbb{C}^{K\times M}$ represents the channel estimation noise independent of the estimated channel, which has i.i.d entries with zero mean and unit variance, and $\tau_D \in [0,1]$ reflects the estimation accuracy of the channel coefficient matrix $\hat{\mathbf{g}}_m$. Under the condition of $\tau_D = 0$, we have perfect CSI estimation. Under the condition of $\tau_D = 1$, the CSI is completely unknown. The CSI may be imperfect due to the estimation errors, limited CSI feedback quantization, delays, etc. On the other hand, the Quality of Experience (QoE) of computation heavily relies on the wireless fading channel conditions since task offloading requires effective wireless transmission. Thus, we consider imperfect CSI for task offloading. Given the imperfect CSI of receiver, the precoding matrix of the CAN is similar to [9], [10], which can be expressed as

$$\hat{\mathbf{g}}^*_m = \hat{\mathbf{g}}^\dagger_m, \tag{22}$$

where $\hat{\mathbf{g}}^\dagger_m = (\hat{\mathbf{g}}^{\mathrm{H}}_m \hat{\mathbf{g}}_m)^{-1}\hat{\mathbf{g}}^{\mathrm{H}}_m$.

The multiplied symbol of $\hat{\mathbf{g}}^*_{m,k}\mathbf{y}_m$ is then delivered to CPU via backhaul link using high-throughput optical fiber for signal detection. Then, the received data at CPU is aggregated, which

can be expressed as

$$\begin{aligned}\mathbf{r}_{\mathrm{CPU}} =& \sum_{m=1}^{M} \hat{\mathbf{g}}_m^* \mathbf{y}_m \\ =& \sqrt{\rho q_k} \sum_{m=1}^{M} \mathbf{g}_m^{\dagger} \mathbf{g}_m \mathbf{s} + \sum_{m=1}^{M} \mathbf{g}_m^{\dagger} \mathbf{n}_m \\ =& \sqrt{\rho q_k} \sum_{m=1}^{M} (\hat{\mathbf{g}}_m^{\mathrm{H}} \hat{\mathbf{g}}_m)^{-1} \hat{\mathbf{g}}_m^{\mathrm{H}} \left(\sqrt{1-\tau_D^2} \hat{\mathbf{g}}_m + \tau_D \mathbf{\Omega}_D \right) \mathbf{s} \\ &+ \sum_{m=1}^{M} (\hat{\mathbf{g}}_m^{\mathrm{H}} \hat{\mathbf{g}}_m)^{-1} \hat{\mathbf{g}}_m^{\mathrm{H}} \mathbf{n}_m \\ =& \sqrt{\rho q_k (1-\tau_D^2)} \sum_{m=1}^{M} \mathbf{s} + \tau_D \sqrt{\rho q_k} \sum_{m=1}^{M} (\hat{\mathbf{g}}_m^{\mathrm{H}} \hat{\mathbf{g}}_m)^{-1} \hat{\mathbf{g}}_m^{\mathrm{H}} \mathbf{\Omega}_D \mathbf{s} \\ &+ \sum_{m=1}^{M} (\hat{\mathbf{g}}_m^{\mathrm{H}} \hat{\mathbf{g}}_m)^{-1} \hat{\mathbf{g}}_m^{\mathrm{H}} \mathbf{n}_m \end{aligned} \tag{23}$$

[1]

Denote r_k as the aggregated received signal for TN k in (23), which can be expressed as

$$\begin{aligned} r_k =& \sqrt{\rho(1-\tau_D^2)} \sum_{m=1}^{M} \sqrt{q_k} s_k + \\ & \tau_D \sqrt{\rho q_k} \sum_{m=1}^{M} (\hat{\mathbf{g}}_{m,k}^{\mathrm{H}} \hat{\mathbf{g}}_{m,k})^{-1} \hat{\mathbf{g}}_{m,k}^{\mathrm{H}} \mathbf{\Omega}_{D,k} s_k + \\ & \sum_{m=1}^{M} (\hat{\mathbf{g}}_{m,k}^{\mathrm{H}} \hat{\mathbf{g}}_{m,k})^{-1} \hat{\mathbf{g}}_{m,k}^{\mathrm{H}} n_{m,k} \end{aligned} \tag{24}$$

Consider the worst-case of the uncorrelated Gaussian noise, we can obtain the received SINR of the transmitted signal in (24) [36], which is shown as (25) in the bottom of this page. As cell-free massive MIMO systems support a very large number of CANs serving a small number of FANs, we characterize the asymptotic property of the received SINR in (25) with the cell-free massive MIMO setting in the following theorem. By this theorem, we can solve the task allocation subproblem.

Theorem 1. *As the number of CANs goes to infinity, i.e., $M \to \infty$, the received SINR in* (25) *is capable of being asymptotically expressed as*

$$\gamma_k = \frac{q_k \rho (1-\tau_D^2)}{\frac{\sigma^2}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \lambda_{k,m}^{-1}}. \tag{26}$$

Proof: The proof is given in Appendix A. ∎

Based on Theorem 1, the received SINR for TN k only related to transmit power from the kth FAN for very large M.

B. Bandwidth Allocation

As second-order derivative of the objective function is strictly negative, we have that T_{total} is a convex function of $\boldsymbol{\eta}$, so we obtain that optimization problem (16) is a convex optimization problem. Then, we refer to the Lagrangian multiplier method to solve problem (16). We denote κ, ν and ξ as the Lagrange multipliers associated with each constraint in problem (16). Consequently, we have the Lagrangian function as

$$\begin{aligned} L(\boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\nu}, \boldsymbol{\xi}) = \Omega_{\mathrm{total}}(\boldsymbol{\eta}) + \sum_{k=1}^{K} \kappa_k \eta_k + \sum_{k=1}^{K} \nu_k (1-\eta_k) \\ + \sum_{k=1}^{K} \xi_k (\sum_{k=1}^{K} \eta_k - B). \end{aligned} \tag{27}$$

Then, we have

$$\begin{aligned} \frac{\partial L(\boldsymbol{\eta}, \boldsymbol{\kappa}, \boldsymbol{\nu}, \boldsymbol{\xi})}{\partial \eta_k} =& \left(\frac{-\mu_k p_{b,m} l_i}{\eta_k^2 B \log_2 \left(1 + \frac{p_{b,m} d_{b,m} |h_{b,m}|^2}{\sigma_m^2}\right)} - \right. \\ & \left. \frac{(1-\mu_k) l_k}{\eta_k^2 B \log_2 \left(1 + \frac{p_{b,m} d_{b,m} |h_{b,m}|^2}{\sigma_m^2}\right)} \right) + \kappa_k \\ & - \nu_k + \xi_k. \end{aligned} \tag{28}$$

Next, the Karush-Kuhn-Tucker (KKT) conditions can be expressed as

$$\frac{\partial L(\boldsymbol{\eta}^*, \boldsymbol{\kappa}, \boldsymbol{\nu}, \boldsymbol{\xi})}{\partial \eta_k^*} = 0, \quad \forall k \in \mathcal{K}, \tag{29}$$

$$\kappa_k \eta_k^* = 0, \quad \nu_k (1-\eta_k^*) = 0, \quad \forall k \in \mathcal{K}, \tag{30}$$

$$\xi_k (\sum_{k=1}^{K} \eta_k^* - B) = 0, \quad \forall k \in \mathcal{K}, \tag{31}$$

$$\sum_{k=1}^{K} \eta_k^* = B, \quad 0 \le \eta_k^* \le 1, \quad \forall k \in \mathcal{K}. \tag{32}$$

Based on (28) and (29), we arrive at

$$\begin{aligned} \xi_k =& \frac{1}{(\eta_k^*)^2} \left(\frac{(1-\mu_k) l_k}{B \log_2 \left(1 + \frac{p_b d_{b,m} |h_{b,m}|^2}{\sigma_m^2}\right)} + \right. \\ & \left. \frac{\mu_k p_b l_k}{B \log_2 \left(1 + \frac{p_b d_{b,m} |h_{b,m}|^2}{\sigma_m^2}\right)} \right) - \kappa_k + \nu_k. \end{aligned} \tag{33}$$

As a result, the optimal solution can be summarized as follows:

- Under the condition that

$$\begin{aligned} \xi_k \ge& \frac{1}{(\eta_k^*)^2} \left(\frac{(1-\mu_k) l_k}{B \log_2 \left(1 + \frac{p_b d_{b,m} |h_{b,m}|^2}{\sigma_m^2}\right)} + \right. \\ & \left. \frac{\mu_k p_b l_k}{B \log_2 \left(1 + \frac{p_b d_{b,m} |h_{b,m}|^2}{\sigma_m^2}\right)} \right). \end{aligned} \tag{34}$$

we have $\eta_k^* = 0$, $\nu_k = 0$ and $\kappa_k \ge 0$ based on (30).

[1] Similar to [20], [36], the CANs are transmitted with the CPU through perfect communication links. These perfect communication links might be established through optical fiber between the CANs and CPU. Additionally, the communication links between CANs and CPU can be realized by copperbased backhaul links, which is capable of providing a capacity of 750 Mbits/s with the maximum range of 1.5 km according to [38].

- Under the condition that

$$\xi_k \leq \frac{1}{(\eta_k^*)^2}\left(\frac{(1-\mu_k)l_k}{B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}+\frac{\mu_k p_b l_k}{B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}\right). \tag{35}$$

we have $\eta_k^*=1$, $\kappa_k=0$ and $\nu_k\geq 0$ based on (30).
- If $0<\eta_k^*<1$, we have $\nu_k=\kappa_k=0$ based on (30). Thus, we have

$$\xi_k = \frac{1}{(\eta_k^*)^2}\left(\frac{(1-\mu_k)l_k}{B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}+\frac{\mu_k p_b l_k}{B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}\right). \tag{36}$$

Then, the optimal solution can be derived as $\eta_k^*=\sqrt{\frac{(1-\mu_k)l_k}{\xi_k B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}+\frac{\mu_k p_{b,m} l_k}{\xi_k B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}}$, $\forall k\in\mathcal{K}$.

Since $0<\eta_k<1$, $\forall k$, the optimal bandwidth allocation results $\boldsymbol{\eta}^*$ of Problem (16) can be obtained as (37), which is shown in the bottom of the next page.

C. Optimizing Task Allocation

Since problem (16) is a mixed integer programming problem with respect to $\boldsymbol{\alpha}$, to obtain its globally optimal solution is NP-hard. Under the circumstances, the partially dualized method of [39] can be adopted to solve such problem. Then, we denote $\boldsymbol{\psi}=[\psi_1,\cdots,\psi_K]$ as the dual variable for the constraint (16d). Consequently, the Lagrangian function can be expressed as

$$\begin{aligned}
&F(\boldsymbol{\alpha},\boldsymbol{\psi}) = \Omega_{\text{total}}(\boldsymbol{\alpha})+\sum_{k\in\mathcal{K}}\psi_k\left(\sum_{k\in\mathcal{K}}\alpha_k-K\right)\\
&=\sum_{k\in\mathcal{K}}\left\{\frac{(1-\mu_k)l_k}{\eta_k B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}+\frac{C_k(1-\mu_k)(1-\alpha_k)l_k}{f_k^F}\right.\\
&\quad+\frac{(1-\mu_k)\alpha_k l_k}{B\log_2(1+\gamma_k)}+\frac{C_k(1-\mu_k)\alpha_k l_k}{f_k^C}+\\
&\quad\frac{\mu_k p_b l_k}{\eta_i B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}+C_0\mu_k(1-\alpha_k)l_k P_{\text{CN}}\\
&\quad\left.+\frac{\mu_k q_k\alpha_k l_k}{B\log_2(1+\gamma_k)}\right\}+\psi\left(\sum_{k\in\mathcal{K}}\alpha_k-K\right)\\
&=\sum_{k\in\mathcal{K}}\left(\frac{(1-\mu_k)l_k}{B\log_2(1+\gamma_k)}+\frac{(1-\mu_k)C_k l_k}{f_k^C}-\frac{(1-\mu_k)C_k l_k}{f_k^F}+\right.\\
&\quad\left.\frac{\mu_k q_k l_k}{B\log_2(1+\gamma_k)}-C_0\mu_k l_k P_{\text{CN}}+\psi\right)\alpha_k+K\psi\\
&\quad+\sum_{k\in\mathcal{K}}\left(\frac{(1-\mu_k)l_k}{\eta_k B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}+\frac{C_k l_k}{f_k^F}\right.\\
&\quad\left.+\frac{\mu_k p_b l_k}{\eta_i B\log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}\right).
\end{aligned}\tag{38}$$

In this case, we can obtain the the partially dualized problem of the original Problem (16), which is given by

$$\begin{aligned}
f(\boldsymbol{\alpha},\boldsymbol{\psi}) = \min_{\alpha_j}\quad & F(\{\alpha_j\},\boldsymbol{\psi}),\\
\text{s.t.}\quad \alpha_j \;&\in\; \{0,1\},\forall j\in\mathcal{K},\\
\textstyle\sum_{j=1}^K\alpha_j \;&\leq\; K.
\end{aligned}\tag{39}$$

Then, the analytical task allocation results can be explicitly obtained for fixed dual variable ψ_k, which is shown as (40) in the bottom of the next page. As k^* is not unique, the tasks will be computed remotely at CAN with $k^*=1$. The optimal dual variable ψ_k^* can be obtained by applying the subgradient approach of [39] or the coordinate descent approach [40]. According to the above results, we have a theorem shown in the following.

Theorem 2. *For binary task offloading in cell-free massive MIMO systems, when the number of CANs tends to infinity, i.e., $M\to\infty$, all the tasks will be computed remotely at CAN for the delay tolerant services, i.e., $\mu_k=1$, $\forall k$. On the other hand, only the tasks from node i^* will be computed remotely at CAN, where FAN i^* has the lest computational capacity. The other tasks will be computed locally at the FAN for the delay sensitive services, i.e., $\mu_k=0$, $\forall k$.*

Proof: The proof is given in Appendix B. ■

$$\gamma_k=\frac{\rho(1-\tau_D^2)|\sum_{m=1}^M\sqrt{q_k}|^2}{\rho\tau_D^2 q_k|\sum_{m=1}^M(\hat{\mathbf{g}}_{m,k}^{\text{H}}\hat{\mathbf{g}}_{m,k})^{-1}\hat{\mathbf{g}}_{m,k}^{\text{H}}\boldsymbol{\Omega}_{D,k}s_k|^2+|\sum_{m=1}^M(\hat{\mathbf{g}}_{m,k}^{\text{H}}\hat{\mathbf{g}}_{m,k})^{-1}\hat{\mathbf{g}}_{m,k}^{\text{H}}n_{m,k}|^2} \tag{25}$$

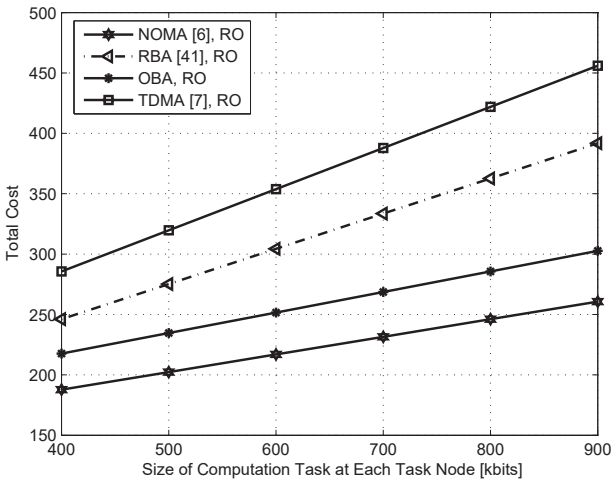

Fig. 2. Total task computation cost versus the task size, for $K = 8$.

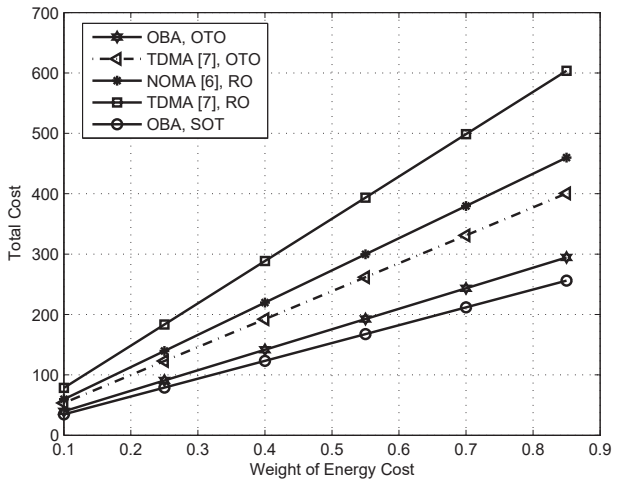

Fig. 3. Total cost versus the weight of energy cost μ_k, $\forall k$, for $K = 8$.

V. Performance Evaluation

In this section, our simulation results characterize the performance of bandwidth allocation and task allocation for the proposed delay-aware cell-free massive MIMO-aided multi-tier computing systems compared to several baseline schemes.

A. System Parameters

To verify the accuracy of the analysis results, we run simulations over $10,000$ channel realizations to obtain the averaging results. Additionally, the channel coefficient samples are generated at a period of 0.005ms during the simulation. Unless otherwise noted, most simulations follow the following scenario. The cell-free massive MIMO systems have 25 FANs with sufficient computing resources. The computational capacity of each CN is selected from the set $\{0.2, 0.3, \cdots, 0.8\}$ GHz randomly and will be fixed. The local computing energy per computation cycle z_i follows a uniform distribution in the range of $(0, 20 \times 10^{-11})$ J/cycle. For the computational task, we consider the service similar to that in [8], where any task k_s has a size of $l_k = 500$ KB, $\forall k \in \mathcal{S}$, and the required computation cycles per bit follows a uniform distribution in the range of $[500, 1500]$ cycles/bit.

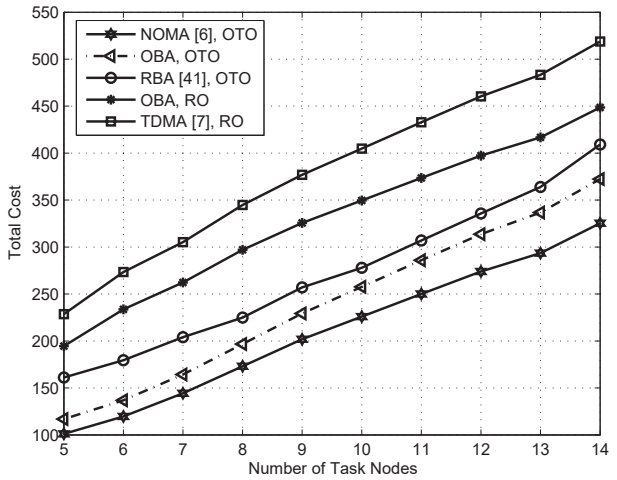

Fig. 4. Total cost versus the number of TNs K.

B. Performance bandwidth allocation

In Fig. 2, we show the total cost under different sizes of the computational task at each TN, where we compare our opti-

$$\eta_k^* = \sqrt{\frac{(1-\mu_k)l_k}{\xi_k B \log_2\left(1+\frac{p_b d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)} + \frac{\mu_k p_{b,m} l_k}{\xi_k B \log_2\left(1+\frac{p_{b,m} d_{b,m}|h_{b,m}|^2}{\sigma_m^2}\right)}}, \quad \forall k. \tag{37}$$

$$\alpha_k^* = \begin{cases} 1, & \text{if } k^* = \arg\min\limits_{k' \in \mathcal{K}} \left(\frac{(1-\mu_{k'})l_{k'}}{B\log_2(1+\gamma_{k'})} + \frac{(1-\mu_{k'})C_{k'}L_{k',s}}{f_{k'}^C} - \frac{(1-\mu_{k'})C_{k'}l_{k'}}{f_{k'}^F} + \right. \\ & \left. \frac{\mu_{k'}q_{k'}l_{k'}}{B\log_2(1+\gamma_{k'})} - C_0\mu_{k'}l_{k'}P_{\text{CN}} + \psi \right), \\ 0, & \text{otherwise}, \end{cases} \tag{40}$$

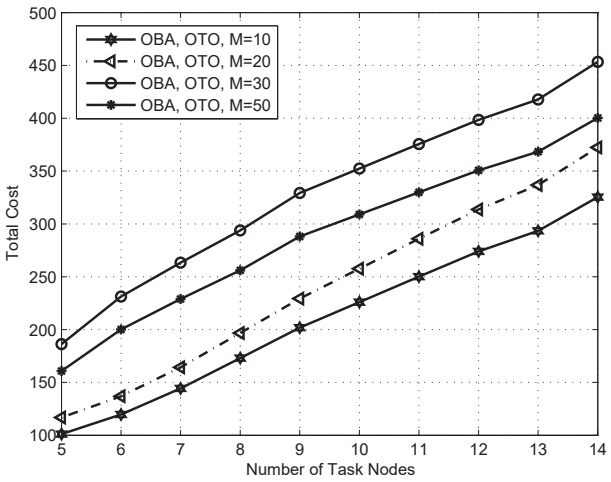

Fig. 5. Total cost versus different number of TNs K for different number of CANs.

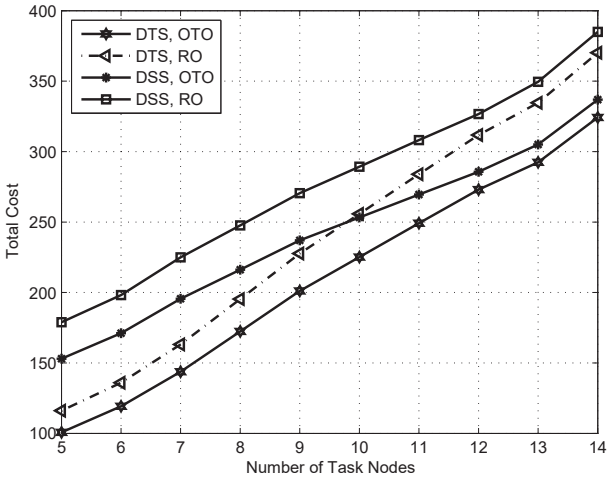

Fig. 6. Total cost versus different number of TNs for different delay requirement services.

mal bandwidth allocation (OBA) strategy to the conventional techniques of time division multiple access (TDMA) [7], non-orthgonal multiple access (NOMA) [6] and random bandwidth allocation (RBA) [41]. Explicitly, in the NOMA strategy, each TN offloads the task with the same time and frequency. It can be observed from this figure that the NOMA performs best in total energy and delay cost compared to TDMA and FDMA. This is because the task offloading latency can be largely reduced by NOMA strategy. However, as the received signals from different TNs are superimposed at the FANs, this strategy has relatively highest computational complexity than TDMA and FDMA. As expected, we can observe from this figure that our OBA strategy always performs best in total cost than those of the TDMA and RBA strategies for different sizes of computational task. Furthermore, we can observe that the NOMA offers better performance in total cost than that of the OBA strategy, since OBA strategy is derived from FDMA. Upon increasing the size of the computational task, there is an another interesting remark, the total cost is increased. If the number of size of the computational task is higher, then the task computation energy consumption will be higher. Additionally, it can be observed from this figure that the OBA strategy offers better performance than TDMA in terms of total energy and delay consumption, implying that it is more beneficial for optimal bandwidth allocation in total energy and delay consumption.

For the comparison of different bandwidth allocation strategies, we also plot the total cost when applying simulated optimal task offloading (SOT), optimized task offloading (O-TO) and random task offloading (RO) strategies. In Fig. 3, we show the total task offloading cost versus the weight of energy consumption μ_k, $\forall k$, for different task offloading strategies. We can observe from this figure that the total cost increases when increasing μ_k, implying that the total energy consumption dominates the total cost. It can be observed that the total task offloading cost of OBA strategy is lower than that of the TDMA under the condition of the same task offloading policy. These observations further verify that the system cost can be reduced by different weights of energy consumption with OBA policy. In summary, OBA strategy makes the task offloading more power efficient without degrading corresponding task offloading rate. As expected, it can be observed from this figure that SOT always performs best in total cost than those of the OTO and RO strategies for different weights of energy consumption. Furthermore, we can observe that the OTO strategy always reduces the total cost regardless of the bandwidth allocation strategies, which confirms the analytical results. In case of energy efficient task offloading, our simulation results further indicate that taking bandwidth allocation into account is necessary, which can realize energy-efficient task offloading for different delay requirements.

C. Performance of task allocation

As shown analytically in Section IV, the cell-free massive MIMO gain of the bandwidth allocation strategy over the task offloading strategy decreases with the number of CAN as well as the received SINR, while it is not affected by these parameters when the number of CAN goes to infinity. In this section, we demonstrate the impact of M, K, and different requirements of the computational tasks on the total cost by means of simulations.

Fig. 4 illustrates the total cost versus different number of TNs for different bandwidth allocation and task offloading policies. We compare our strategy to traditional systems operating with RO, such that RO serves as a benchmark. As expected, we can observed from this figure that our OTO strategy always performs better in total energy and delay cost than that of the RO policy with different number of TNs. Additionally, it can observed that the OTO scheme relying on the optimal bandwidth allocation strategy can always achieve better performance than that of the traditional offloading scheme. Furthermore, as expected, optimal bandwidth allocation scheme offers much better performance than TDMA and

RBA utilizing the same task allocation scheme. It should be noted that even if the OBA scheme is adopted, OTO performs better by the optimal task allocation scheme with the aid of cell-free massive MIMO. This is due to the fact that the optimal task allocation scheme increases the task offloading throughput as well as decreases its energy cost of the cell-free massive MIMO systems. On the other hand, it can be observed that the total cost is increased by increasing the number of TNs. It should be noted that if the number of TNs is larger, then the energy cost and total computational delay of the task offloading cell-free massive MIMO systems will be larger.

Fig. 5 illustrates the performance of the total cost in terms of both different number of TNs and different number of CANs. As expected, it can be observed that the OBA scheme using OTO strategy can reduce total cost. Meanwhile, the total cost is increased by increasing the number of TNs. This result implies that the total energy consumption and computational latency are increased. Furthermore, we can observe that the larger the number of CAN is, the smaller the total cost is. This is because the larger the number of CAN, the higher the SINR at the CPU, thus reducing both the total task offloading delay and energy cost. As the optimal task allocation strategy can realize energy-efficient task transmission, the larger the number of CAN utlizing the optimal task allocation strategy can always perform better in total cost, which confirms our analysis results. It should be noted that these results indicate that the total cost is decreased by placing more CANs.

Next, Fig. 6 plots the total cost versus different number of TNs K for different delay requirement services. It can be observed that the total cost is significantly increased upon increasing K and the delay tolerant services (DTS) with OTO always performs best in total cost, which is due to the fact that the tasks are allocated to the CPU with cloud computation for DTS with OTO. Additionally, the OTO scheme associated with optimal task allocation scheme always offers better performance than the RO scheme. This is because the optimal task allocation scheme can achieve the minimal delay and energy consumption for task transmission. We can also observe that for different number of TNs, delay sensitive services (DSS) always have much larger computational latency and energy consumption than DTS for the same task offloading strategy. Since the DSS has much more stringent delay requirement, result in larger energy consumption and total cost, which confirms our analysis results. We have another interesting remark from this figure, there is a crossing point by increasing the number of TNs between the performances of the DSS and the DTS for different task offloading strategies. According to these simulation results, when K is small, the DTS with RO strategy has smaller total cost than DSS with OTO scheme. On the other hand, when K is large, the DTS with RO strategy has larger total cost than DSS with OTO scheme. We hold the opinion that the optimal task offloading strategy has a great influence on the total cost when K is large. In this case, the OTO will always performs better regardless of the different delay requirements of the tasks.

VI. Conclusions

In this paper, we proposed a cell-free massive MIMO-assisted multi-tier computing systems and investigated the bandwidth allocation and task offloading, where the intensive tasks from TNs can be offloaded to nearby FAN, and to the CPU constituted by the nearby CAN via the cell-free massive MIMO. We formulated a total cost minimization problem in terms of energy consumption and computational latency, while considering realistic heterogenous delay requirements of the tasks. Since we consider binary task offloading, the resulting non-convex bandwidth and task allocations problem can be solved by decoupling the original problem. As the bandwidth allocation problem is a convex optimization problem, we first obtained the bandwidth allocation solution under a given task allocation strategy, followed by conceiving the traditional convex optimization method to determine the bandwidth allocation result. According to the bandwidth allocation solution, the Lagrange partial relaxation method has been used for formulating the Lagrange dual problem by relaxing the binary constraint to determine the task offloading result. The simulation results demonstrate that the proposed strategy always performs best with the benchmark strategies. Meanwhile, based on the received SINR obtained by $M \to \infty$, the cost optimal task offloading strategy can be chosen for heterogeneous delay requirements of the computational tasks in our proposed cell-free massive MIMO-assisted multi-tier computing systems. Future work is in progress to consider more general case that arbitrary FANs are deployed for each TN to assist task offloading in the proposed framework.

Appendix A
Proof of Theorem 1

According to Tchebyshevs theorem [42], we have

$$\lim_{N\to\infty}\frac{1}{N}(X_1+X_2+\ldots+X_N)=\frac{1}{N}(\mathsf{E}(X_1)+\mathsf{E}(X_2)+\ldots+\mathsf{E}(X_N)), \tag{41}$$

where X_1, X_2, …, X_N are N independent random variables, $\mathsf{E}(X_i)$ denotes the expectation of X_i, $\forall i$. Then, regarding the second term on the right hand side in (24), we have

$$\lim_{M\to\infty}\frac{1}{M}\tau_D\sqrt{\rho q_k}\sum_{m=1}^{M}(\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}\hat{\mathbf{g}}_{m,k})^{-1}\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}\mathbf{\Omega}_{D,k}s_k=0. \tag{42}$$

By adopting the eigenvalue/eigenvector decomposition of $\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}\hat{\mathbf{g}}_{m,k}$, we obtain

$$\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}\hat{\mathbf{g}}_{m,k}=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{H}}, \tag{43}$$

where $\mathbf{\Lambda}=\mathrm{diag}\{\lambda_1,\cdots,\lambda_{\mathrm{K}}\}$ and $\mathbf{Q}$ respectively denotes the nonnegative diagonal eigenvalue matrix and the unitary eigenvector matrix, respectively. Thus, regarding the third term

on the right hand side in (24), we have

$$\begin{aligned}&\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}(\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}\hat{\mathbf{g}}_{m,k})^{-1}\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}n_{m,k}\\&=\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{H}})^{-1}\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}n_{m,k}\\&=\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}(\mathbf{\Lambda})^{-1}\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}n_{m,k}\end{aligned}\tag{44}$$

Then, we have

$$\begin{aligned}&\lim_{M\to\infty}\frac{1}{M}|\sum_{m=1}^{M}(\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}\hat{\mathbf{g}}_{m,k})^{-1}\hat{\mathbf{g}}_{m,k}^{\mathrm{H}}n_{m,k}|^2\\=&\lim_{M\to\infty}\frac{\sigma^2}{M}\sum_{m=1}^{M}\mathrm{tr}\left((\hat{\mathbf{g}}_{\mathrm{m,k}}^{\mathrm{H}}\hat{\mathbf{g}}_{\mathrm{m,k}})^{-1}\hat{\mathbf{g}}_{\mathrm{m,k}}^{\mathrm{H}}\hat{\mathbf{g}}_{\mathrm{m,k}}(\hat{\mathbf{g}}_{\mathrm{m,k}}^{\mathrm{H}}\hat{\mathbf{g}}_{\mathrm{m,k}})^{-1}\right)_{\mathrm{k,k}}\\=&\frac{\sigma^2}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\lambda_{k,m}^{-1}.\end{aligned}\tag{45}$$

Based on (42) and (45), we have

$$\lim_{M\to\infty}\gamma_k=\frac{q_k\rho(1-\tau_D^2)}{\frac{\sigma^2}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}\lambda_{k,m}^{-1}}.\tag{46}$$

APPENDIX B
PROOF OF THEOREM 2

While all tasks are from delay tolerant services, the energy consumption will dominate the total cost. Then, $\mu_k=1$, $\forall k$. Substitute $\mu_k=1$ into (40), we have

$$\alpha_k^*=\begin{cases}1, & \text{if } k^*=\arg\min\limits_{k'\in\mathcal{K}}(\frac{ql}{B\log_2(1+\gamma_{k'})}-C_0lP_{\mathrm{CN}}+\psi),\\0, & \text{otherwise.}\end{cases}\tag{47}$$

As a result, $\alpha_k^*=1$, $\forall k$. Therefore, all the tasks will be computed remotely at CAN.

On the other hand, while all the tasks are from delay sensitive services, the total delay cost will dominate the total cost. Then, $\mu_k=0$, $\forall k$. Substitute $\mu_k=1$ into (40), we have

$$\alpha_k^*=\begin{cases}1, & \text{if } k^*=\arg\min\limits_{k'\in\mathcal{K}}(\frac{l}{B\log_2(1+\gamma_{k'})}+\frac{C_{k'}l}{f_{k'}^C}-\frac{C_{k'}l}{f_{k'}^F}+\psi),\\0, & \text{otherwise.}\end{cases}\tag{48}$$

Then, only the tasks from node i^* will be computed remotely at CAN, where $i^*=\arg\min\limits_{i\in\mathcal{K}}(\frac{l_i}{B\log_2(1+\gamma_i)}-\frac{C_il_i}{f_i^F})$, i.e., i^* is from the smallest f_k^F, $\forall k$. This indicates that the FAN with the lest computational capacity will offload the tasks to the CAN, while the other tasks will be computed locally at the FAN.

REFERENCES

[1] K. Wang and W. Chen, "Energy-efficient communications in MIMO systems based on adaptive packets and congestion control with delay constraints," *IEEE Trans. Wireless Commun.*, vol. 14, no.4, pp. 2169–2179, Apr. 2015.

[2] C. She, Z. Chen, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Improving network availability of ultra-reliable and low-latency communications with multi-connectivity," *IEEE Trans. Commun.*, vol. 66, no.11, pp. 5482–5496, Jun. 2018.

[3] S. Verma, Y. Kawamoto, Z. M. Fadlullah, H. Nishiyama, and N. Kato, "A survey on network methodologies for real-time analytics of massive IoT data and open research issues," *IEEE Commun. Surveys Tuts.*, vol. 19, no.3, pp. 1457–1477, Apr. 2017.

[4] C. She, C. Liu, T. Q. S. Quek, C. Yang, and Y. Li, "Ultra-reliable and low-latency communications in unmanned aerial vehicle communication systems," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3768–3781, May 2019.

[5] D. V. Huynh, S. R. Khosravirad, A. Masaracchia, O. A. Dobre, and T. Q. Duong, "Edge intelligence-based ultra-reliable and low-latency communications for digital twin-enabled metaverse," *IEEE Commun. Lett.*, vol. 11, no. 8, pp. 1733–1737, Aug. 2022.

[6] K. Wang, Y. Zhou, Z. Liu, Z. Shao, X. Luo, and Y. Yang, "Online task scheduling and resource allocation for intelligent NOMA-based industrial Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 38, no.5, pp. 803–815, May 2020.

[7] Y. Yang, K. Wang, G. Zhang, X. Chen, X. Luo, and M.-T. Zhou, "MEETS: Maximal energy efficient task scheduling in homogeneous fog networks," *IEEE Internet Things J.*, vol. 5, pp. 4076–4087, Oct. 2018.

[8] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Networking*, vol. 24, pp. 2795–2808, Oct. 2016.

[9] K. Wang, Y. Zhou, J. Li, S. L, W. Chen, and L. Hanzo, "Energy-efficient task offloading in massive MIMO-aided multi-pair fog-computing networks," *IEEE Trans. Commun.*, vol. 69, no.4, pp. 2123–2137, Apr. 2021.

[10] K. Wang, W. Chen, J. Li, Y. Yang, and L. Hanzo, "Joint task offloading and caching for massive MIMO-aided multi-tier computing networks," *IEEE Trans. Commun.*, vol. 70, no.3, pp. 1820–1833, Mar. 2022.

[11] K. Wang, Y. Zhou, Q. Wu, W. Chen, and Y. Yang, "Task offloading in hybrid intelligent reflecting surface and massive MIMO relay networks," *IEEE Trans. Wireless Commun.*, vol. 21, no.6, pp. 3648–3663, Jun. 2022.

[12] L. D. Nguyen, T. Q. Duong, H. Q. Ngo, and K. Tourki, "Energy efficiency in cell-free massive MIMO with zero-forcing precoding design," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1871–1874, Apr. 2017.

[13] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, M. Debbah, and P. Xiao, "Maxcmin rate of cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. on Commun*, vol. 67, no.10, pp. 6796–6815, Oct. 2019.

[14] K. Wang, J. Jin, Y. Yang, T. Zhang, A. Nallanathan, C. Tellambura, and B. Jabbari, "Task offloading with multi-tier computing resources in next generation wireless networks," May 2022. [Online]. Available: https://arxiv.org/pdf/2205.13866.pdf.

[15] K. Wang, Y. Tan, Z. Shao, S. Ci, and Y. Yang, "Learning-based task offloading for delay-sensitive applications in dynamic fog networks," *IEEE Trans. Veh. Tech.*, vol. 68, no.11, pp. 11399–11403, Nov. 2019.

[16] T. Bai, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Latency minimization for intelligent reflecting surface aided mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 38, no.11, pp. 2666–2682, Nov. 2020.

[17] T. Do-Duy, D. V. Huynh, O. A. Dobre, B. Canberk, and T. Q. Duong, "Digital twin-aided intelligent offloading with edge selection in mobile edge computing," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 806–810, Jan. 2022.

[18] Y. Yang, Z. Liu, X. Yang, K. Wang, X. Hong, and X. Ge, "POMT: Paired offloading of multiple tasks in heterogeneous fog networks," *IEEE Internet Things J.*, vol. 6, pp. 8658–8669, Oct. 2019.

[19] Z. Liu, Y. Yang, K. Wang, Z. Shao, and J. Zhang, "POST: Parallel offloading of splittable tasks in heterogeneous fog networks," *IEEE Internet Things J.*, vol. 7, pp. 3170–3183, Apr. 2020.

[20] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no.7, pp. 4445–4459, Jul. 2017.

[21] S. Mukherjee and J. Lee, "Edge computing-enabled cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no.4, pp. 2884–2899, Apr. 2020.

[22] M. Ke, Z. Gao, Y. Wu, X. Gao, and K.-K. Wong, "Massive access in cell-free massive MIMO-based internet of things: Cloud computing and edge computing paradigms," *IEEE J. Sel. Areas Commun.*, vol. 39, no.3, pp. 756–772, Mar. 2021.

[23] G. Zhang, W. Zhang, Y. Cao, D. Li, and L. Wang, "Energy-delay tradeoff for dynamic offloading in mobile-edge computing system with energy harvesting devices," *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 4642–465, Oct. 2018.

[24] A. Bozorgchenani, F. Mashhadi, D. Tarchi, and S. A. S. Monroy, "Multi-objective computation sharing in energy and delay constrained

mobile edge computing environments," *IEEE Transactions on Mobile Computing*, vol. 20, pp. 2992–3005, Oct. 2021.
[25] Z. Li, N. Zhu, D. Wu, H. Wang, and R. Wang, "Energy-efficient mobile edge computing under delay constraints," *IEEE Transactions on Green Communications and Networking*, vol. 6, pp. 776–786, Jun. 2022.
[26] Y. Deng, Z. Chen, X. Yao, S. Hassan, and A. M. A. Ibrahim, "Parallel offloading in green and sustainable mobile edge computing for delay-constrained IoT system," *IEEE Trans. Veh. Tech.*, vol. 68, pp. 12202–12214, Dec. 2019.
[27] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun.*, pp. 201–205, Jun. 2015.
[28] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 16, no.9, pp. 5994–6009, Sept. 2017.
[29] O. Y. Bursalioglu, G. Caire, R. K. Mungara, H. C. Papadopoulos, and C. Wang, "Fog massive MIMO: A user-centric seamless hot-spot architecture," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 559–574, Jan. 2019.
[30] K. Wang, W. Chen, J. Li, and B. Vucetic, "Green MU-MIMO/SIMO switching for heterogeneous delay-aware services with constellation optimization," *IEEE Trans. Commun.*, vol. 64, no.5, pp. 1984–1995, May 2016.
[31] S. Zhao, Y. Yang, Z. Shao, X. Yang, H. Qian, and C.-X. Wang, "FEMOS: Fog-enabled multitier operations scheduling in dynamic wireless networks," *IEEE Internet Things J.*, vol. 5, no.2, pp. 1169–1183, Apr. 2018.
[32] X. Lyu, H. Tian, W. Ni, Y. Zhang, P. Zhang, and R. P. Liu, "Energy-efficient admission of delay-sensitive tasks for mobile edge computing," *IEEE Trans. Commun.*, vol. 66, no.6, pp. 2603–2616, Jun. 2018.
[33] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.
[34] 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, Aug. 2017.
[35] X. Xiao, X. Tao, and J. Lu, "QoS-aware energy-efficient radio resource scheduling in multi-user OFDMA system," *IEEE Commun. Lett.*, vol. 17, no.1, pp. 75–78, Jan. 2013.
[36] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no.3, pp. 1834–1850, Mar. 2017.
[37] B. Nosrat-Makouei, J. G. Andrews, and J. Robert W. Heath, "MIMO interference alignment over correlated channels with imperfect CSIT," *IEEE Trans. Signal Process.*, vol. 59, no.6, pp. 2783–2794, Jun. 2011.
[38] M. A. Imran, S. A. R. Zaidi, and Z. Shakir, *Access, Fronthaul and Backhaul Networks for 5G & Beyond*. Edison, NJ, USA: Institution of Engineering and Technology, 2017.
[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[40] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no.6, pp. 1100–1113, Jun. 2014.
[41] Z. Song, Y. Liu, and X. Sun, "Joint task offloading and resource allocation for NOMA-enabled multi-access mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 69, no.3, pp. 1548–1564, Mar. 2021.
[42] H. Cramer, *Random Variables and Probability Distributions*. Cambridge, U.K.: Cambridge Univ. Press, 1970.