On Ergodic Energy Efficiency of mmWave Heterogeneous Cell-Free Systems with ZF Hybrid Precoders

Jichong Guo, *Member, IEEE,* Dekun Zhang, Chen Cui, *Member, IEEE,* and Xiqing Liu†, *Member, IEEE*

***Abstract*—Energy efficiency (EE) takes a significant place in the upcoming sixth generation (6G) mobile networks. Cell-free systems have demonstrated impressive performance and attracted much attention in academia and industry. Various types and numbers of access points (APs) can be deployed in heterogeneous cell-free systems, operating at millimeter-wave (mmWave) frequencies. But, the instantaneous EE cannot give a full insight into these systems. To tackle this issue, in this paper, we study the ergodic EE of mmWave heterogeneous cell-free systems. Since the derivation of a general closed-form expression for the ergodic EE is intractable, a semi-closed form expression was achieved via discretization, and its effects were investigated. Based on this expression, two specific cases of gain were obtained to describe the EE performance of the mmWave heterogeneous cell-free systems: the first determines the necessity to employ mmWave heterogeneous cell-free systems, and the second guides their deployment. Our simulation results validate the effectiveness of the two gain cases.**

***Index Terms*—Millimeter-wave communications, cell-free systems, heterogeneous systems, energy efficiency.**

I. Introduction

MILLIMETER-WAVE (mmWave) wireless communications can allow us to dramatically improve the transmission rate by exploring the spectra that are not overexploited [1]. However, the double-edged transmission characteristics of mmWave frequency bands, like its path loss, call for new networks and technologies to ensure reliable wireless communication. Among these emerging network frameworks, the cell-free system – a key enabling technology for the sixth generation (6G) mobile networks [2] – has been gaining attraction for its impressive ability to improve system performances, such as spectrum efficiency (SE) and bit error rate (BER).

This work was supported by the National Key Research & Development Program of China under Grant No. 2020YFB1806703. (Corresponding author: Xiqing Liu)

Jichong Guo is with the School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215000, China (e-mail: guojichong@usts.edu.cn).

Dekun Zhang is with the State Key Laboratory of Integrated Services Networks, Xidian University, Xian 710071, China (e-mail: 21013110276@stu.xidian.edu.cn).

Chen Cui is with the National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China (e-mail: chencui@pku.edu.cn).

Xiqing Liu is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: liuxiqing@bupt.edu.cn).

Cell-free systems can be viewed as a modified version of distributed systems, first reported in papers [3], [4]. In these works, all the access points (APs) were connected to a central processing unit (CPU) and served all users over the same time-frequency resources. However, this canonical form leads to impractical implementation, and many effective solutions have been proposed to circumvent this problem, such as the user-centric [5], [6] and scalable cell-free framework [7], [8], and limited fronthaul [9], [10]. Owing to its excellent compatibility, many critical technologies have been imported into the cell-free systems, e.g., non-orthogonal multiple access (NOMA) [11], [12], reconfigurable intelligent surfaces (RIS) [13], [14], and unmanned aerial vehicles (UAVs) [15], [16].

The evolution of communication generations and the diversification of services cause significant heterogeneity for the current wireless communication systems. In the heterogeneous cellular networks, various classes of low power nodes (LPNs) are distributed throughout the macro cell network [17], [18]. Similar to the heterogeneous cellular case, diverse types of wireless APs are randomly distributed in the heterogeneous (also called X-assisted) cell-free networks [19]–[21]. Unlike the traditional cell-free systems, these wireless APs are not mandatory to change real-time information via a CPU. For example, it is easy to cooperate between the macrocell and femtocell base stations (BSs) [19], while intractable with UAVs worked as mobile BSs [21]. The ongoing research on the heterogeneous cell-free system is at a very early stage.

Energy efficiency (EE) is a critical concern in the sustainable development of wireless communications [22], [23]. Over the past few years, a lot of research has been conducted to design energy-efficient schemes and to evaluate EE performances [24]–[26]. High EE is a significant design target in cell-free systems, with continued research providing us with several fruitful results in this direction. Specifically, several representative works are listed in Table I. As shown in Table I, most studies considered instantaneous EE as the design target, while only a few of them have focused on average EE [19], [27]. For example, to best utilize the energy harvested at each AP, Hamdi and Qaraqe investigated the issues of online energy cooperation and management problem [27]. They proposed an efficient online algorithm based on energy prediction to minimize power consumption in a given time period. Kim and Shim studied the maximum EE of mmWave heterogeneous cell-free systems under limited feedback [19]. An energy-efficient dominating path selection algorithm is proposed to achieve more than 80% EE improvement over the conventional

TABLE I
SEVERAL REPRESENTATIVE PAPERS ON EE IN THE CELL-FREE SYSTEMS

Target	Technology	Heterogeneity
Maximum instantaneous EE	Power control [28], [29], AP selection [30], beamforming [14]	No
Maximum instantaneous EE	Dominating path selection [19]	Yes
Minimum power consumption in a period	Power control [27]	No
Maximum SE while meeting power constraint	Beamforming [13]	No

channel state information (CSI) feedback-based schemes.

The main reason for the focus on instantaneous EE in the aforementioned studies is that in traditional communication systems, wireless access is often through BSs with wide coverage, excellent computation capability, and complex structure, which are hard to deploy in real-time. Also, energy-efficient schemes are usually designed on the instantaneous CSI to maximize EE. However, the instantaneous EE is insufficient to characterize the randomness of the EE performance. As an alternative method of analysis, several studies applied the average EE performance in their simulation results.

Development of theory and industry have changed the concept of network with the introduction of the ultra-dense network (UDN) [31]–[33] and cell-free systems [3], [4], [34] then the wireless accesses widely take LPNs, such as micro BSs, pico evolved node Bs (eNBs), Femtocells, relays and APs. The miniaturized and intensive wireless accesses have allowed real-time deployment. This situation necessitates the study of the ergodic EE, instead of instantaneous EE, in communication systems, and to the best of our knowledge, the ergodic EE of heterogeneous cell-free systems is still an open issue. The ergodic EE of a centralized mmWave system with a hybrid pre-coding scheme is studied in our previous work [35]. Following it, here, we discuss the more complicated ergodic EE of the mmWave heterogeneous cell-free systems. The major contributions of this work are three-fold:

- Firstly, we give the instantaneous EE by modeling an mmWave heterogeneous cell-free system in which various types and numbers of APs can be deployed. Based on this expression, a generic ergodic EE with the statistical information is achieved;
- The ergodic EE with the statistical information is quite complicated to calculate. Several relaxing methods are employed to derive a semi-closed ergodic EE, aiming to provide a complete insight into the systems. The effect of the main relaxing method on the ergodic EE is theoretically analyzed;
- Two gain cases are derived to demonstrate the importance of heterogeneous cell-free systems and to guide their deployment. The effectiveness of these two gain cases is verified through simulations.

The remainder of the paper is organized as follows: Section II describes a generic mmWave heterogeneous cell-free system model and its corresponding transmit signals. Based on the system model, Section III formulates a generic instantaneous and ergodic EE, and Section IV takes several simplification steps to get a semi-closed expression. The two gain cases and their analysis is also shown in Section IV. Simulation results are presented in Section V, followed by the conclusions in Section VI.

Notation: a, $\mathbf{a}$, $\mathbf{A}$, $\boldsymbol{\mathcal{A}}$, $\mathbb{R}$ and $\mathbb{C}$ stand for a variable, a column vector, a matrix, a set, and the real and complex fields, respectively. $\mathbf{A}^{\mathrm{H}}$, $\mathbf{A}^{*}$, and $\mathbf{A}^{\mathrm{T}}$ represent the conjugate transpose, conjugate, and transpose of $\mathbf{A}$, respectively, while $\mathbf{I}$ refers to the unit matrix. $|\boldsymbol{\mathcal{A}}|$ is the cardinality of the set $\boldsymbol{\mathcal{A}}$. Denoted by Ones(x,y) is the $x \times y$ matrix with each element being one. $\mathrm{diag}(\cdot)$ is the diagonalizing function, and $\mathrm{angle}(\cdot)$ is the angular function. And $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. $\mathcal{CN}(\cdot,\cdot)$ and $\mathcal{N}(\cdot,\cdot)$ denote a circularly-symmetric complex Gaussian distribution and a real Gaussian distribution, respectively.

II. SYSTEM DESCRIPTION

A generic illustration of the deployable heterogeneous cell-free system is shown in Fig. 1. The type of APs includes both hardware and software aspects. In other words, various types of APs may be equipped with different transmitter structures and take diverse technologies, resulting in heterogeneous gains. Considering that hardware is difficult to adjust, here we mainly focus on the heterogeneous gain achieved by hardware and set each type of APs to take the same hybrid pre-coding algorithm.

The numbers of APs and their types are denoted as M and T, where $T \leq M$. The special case $T = 1$ indicates the traditional isomorphic cell-free systems. $\boldsymbol{\mathcal{S}}_t$ is the tth type set of APs, and it corresponds to a complete configuration: working frequency f_t, number of transmit antennas N_t, number of radio frequency (RF) chains l_t, maximum transmit power P_t and so on. If the mth AP belongs to the tth type set of APs, i.e., $m \in \boldsymbol{\mathcal{S}}_t$, it means that the mth AP is equipped with the configuration of $\boldsymbol{\mathcal{S}}_t$. K users are served in an area through this deployable heterogeneous cell-free system, and each user is equipped with a single receiver antenna. The number of served users meets $K \leq \sum_t |\boldsymbol{\mathcal{S}}_t| l_t$, where the equal sign establishes when all the APs serve different users. APs are assigned to the kth user by rules, which are contained in set $\boldsymbol{\mathcal{A}}_k$ as explained in many previous works [5], [6]. Therefore, the users served by the mth AP are contained in a set $\boldsymbol{\mathcal{A}}_m^{-1}$ such that $|\boldsymbol{\mathcal{A}}_m^{-1}| \leq l_t$. The modulated signal of the kth user is denoted by $d_k \in \mathbb{C}$, satisfying $\mathrm{E}[d_k d_k^*] = 1$. Hence, the total modulated signal at the mth AP is denoted as $\mathbf{d}_m = [\cdots, d_k, \cdots]^{\mathrm{T}} \in \mathbb{C}^{|\boldsymbol{\mathcal{A}}_m^{-1}| \times 1}$.

A hybrid transmit architecture always bridges the gap between the digital and analog architectures, and hence, its pre-coding matrix is generic. The pre-coding matrix of the analog transmit architecture is obtained via the digital pre-coding part of the hybrid transmit architecture as a unit matrix. Hence, the mth AP takes a local hybrid precoder, consisting of an analog part $\mathbf{F}_m \in \mathbb{C}^{N_t \times l_t}$ and a digital part $\mathbf{W}_m \in \mathbb{C}^{l_t \times |\boldsymbol{\mathcal{A}}_m^{-1}|}$.

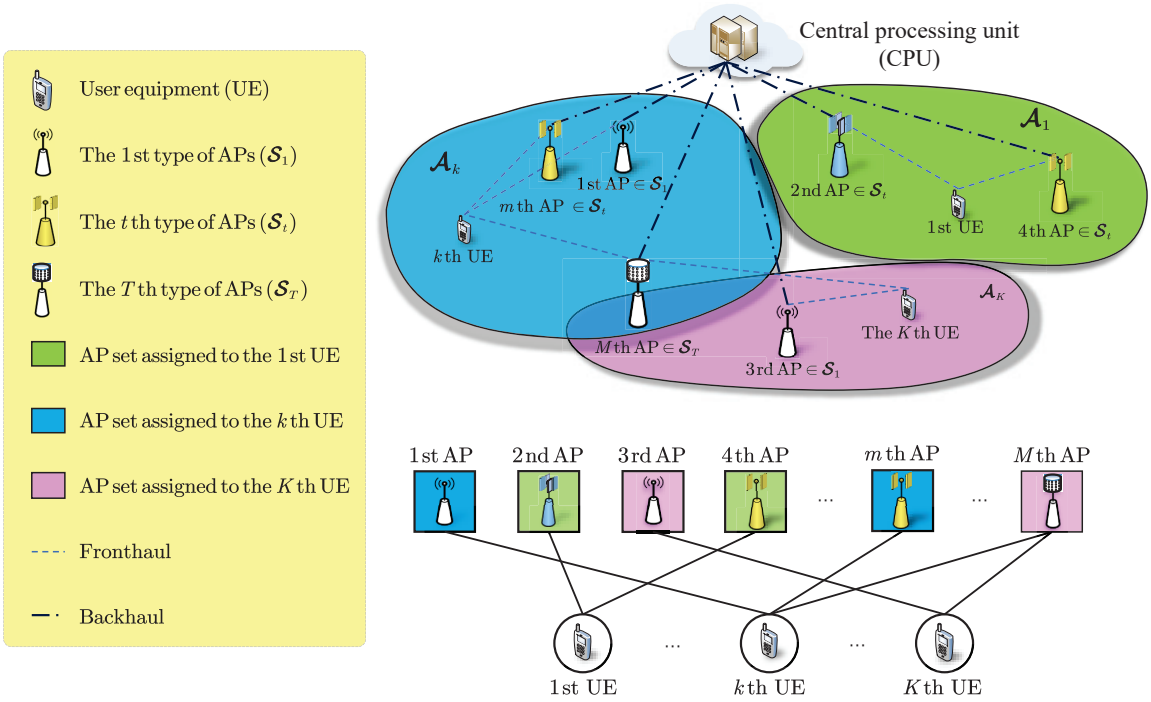


Fig. 1. Illustrations of heterogeneous cell-free systems. $\boldsymbol{\mathcal{S}}_t$ represents the tth type set of APs. As an example, each color corresponds to a set of APs assigned to a user by a rule.

Considering that the antenna gain just affects the value of EE, instead of the trend. In other words, the antenna gain will not affect the theoretical analysis. Then gain of each antenna element is assumed to be normalized. After processed by the power allocation matrix $\mathbf{P}_m \in \mathbb{R}^{|\boldsymbol{\mathcal{A}}_m^{-1}|\times|\boldsymbol{\mathcal{A}}_m^{-1}|}$ and hybrid precoding matrices $\mathbf{F}_m$ and $\mathbf{W}_m$, the transmitted signal at the mth AP is expressed as

$$\mathbf{X}_m = \mathbf{F}_m \mathbf{W}_m \sqrt{\mathbf{P}_m}\mathbf{d}_m. \tag{1}$$

Suppose that fading is constant in a block, and the channel between the mth $m \in \boldsymbol{\mathcal{S}}_t$ AP and the kth user is denoted as $\mathbf{H}_{m,k} \in \mathbb{C}^{N_t\times 1}$. Since the fading is composed of large-scale and small-scale fadings, we have $\mathbf{H}_{m,k} = \mathbf{H}_{m,k}^{\text{small}}\mathbf{I}^{\text{NB}}\sqrt{\mathbf{L}_{m,k}}$, where the diagonal matrix $\mathbf{I}^{\text{NB}}$ indicates the existence of rays. $\mathbf{L}_{m,k} \in \mathbb{C}^{N_{\text{cl},m,k}\times 1}$ is the large-scale fading matrix. Also, $\mathbf{H}_{m,k}^{\text{small}} \in \mathbb{C}^{N_t\times N_{\text{cl},m,k}}$ represents the small-scale fading matrix, in which $N_{\text{cl},m,k}$ is the number of distinguishable paths. Without loss of generality, a column of $\mathbf{H}_{m,k}^{\text{small}}$ is the product of a small-scale fading coefficient and an array steering vector, which is decided by the transmit antenna array and the direction of arrival (DoA) of the ray. In the mmWave communication systems, when the distance between transmitter and receiver is smaller than a threshold value, such as 50 meters, the outage probability of line-of-sight (LoS) path is less than 30% [36]. With the user-centric cell-free systems, we can select the APs owning the LoS path to the kth user, which forms a set $\boldsymbol{\mathcal{A}}_k$. So when the number of APs is large, it is able to ensure the situation that mmWave LoS path exists with a high probability. Therefore, it is reasonable to set one diagonal element of $\mathbf{I}^{\text{NB}}$ corresponding to the LoS path being one.

Generally, $\mathbf{L}_{m,k}$ can be simplified to $\bar{L}_{m,k}\text{Ones}\left(N_{\text{cl},m,k}, 1\right)$ by taking the centroid of all the fadings, with $\bar{L}_{m,k}$ being the mean large-scale fading coefficient. According to the practical measurements in the mmWave communication systems [36], [37], the number of distinguishable paths is small (1 with approximately 48% probability for 28 GHz) and the power difference between the LoS path and the non-line-of-sight (NLoS) path is quite large (at least 5 dB). Then $\bar{L}_{m,k}$ can be approximated by the LoS large-scale fading coefficient, denoted by $L_{m,k}$. This gives us $\mathbf{H}_{m,k} = \sqrt{L_{m,k}}\mathbf{H}_{m,k}^{\text{small}}\text{Ones}\left(N_{\text{cl},m,k}, 1\right) = \sqrt{L_{m,k}}\mathbf{h}_{m,k}$, where $\mathbf{h}_{m,k} \in \mathbb{C}^{N_t\times 1}$. Let us define $\mathbf{h}_m = [\cdots, \mathbf{h}_{m,k}, \cdots] \in \mathbb{C}^{N_t\times|\boldsymbol{\mathcal{A}}_m^{-1}|}$ and $\mathbf{L}_m = [\cdots, L_{m,k}, \cdots] \in \mathbb{C}^{|\boldsymbol{\mathcal{A}}_m^{-1}|\times|\boldsymbol{\mathcal{A}}_m^{-1}|}$ such that the channel matrix of the mth AP becomes $\mathbf{H}_m = \mathbf{h}_m\sqrt{\mathbf{L}_m} = [\cdots, \mathbf{H}_{m,k}, \cdots] \in \mathbb{C}^{N_t\times|\boldsymbol{\mathcal{A}}_m^{-1}|}$.

Particularly, the large-scale fading coefficient $L_{m,k}$ is modeled as [38]

$$\begin{aligned} L_{m,k} = &-20\log_{10}\left(\frac{4\pi}{\lambda}\right) - \\ &10n_0\left[1-b+\frac{bc}{\lambda f_0}\right]\log_{10}\left(r_{m,k}\right) - X_\sigma, \end{aligned} \tag{2}$$

where X_σ represents the logarithmically Gaussian-distributed shadow fading term, with a zero-mean and variance σ^2. The path loss exponent n_0 is always decided by the specific scenarios. In (2), b is an optimization parameter that captures the slope and f_0 is a fixed reference frequency. Besides, c and λ are the speed of light and the carrier wavelength, respectively. The distance of LoS path between the mth AP and the kth user is denoted as $r_{m,k}$.

Then the received signal of the kth user can be written as,

$$y_k = \sum_{m\in\mathcal{A}_k} \mathbf{H}_{m,k}^{\mathrm{T}}\mathbf{F}_m\mathbf{W}_{m,k}\sqrt{P_{m,k}}d_k + z_k + \sum_{k'\neq k}\sum_{m'\in\mathcal{A}_{k'}} \mathbf{H}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}\sqrt{P_{m',k'}}d_{k'}, \tag{3}$$

where $\mathbf{W}_{m,k} \in \mathbb{C}^{l_t\times 1}$ is the kth column of $\mathbf{W}_m$ and $P_{m,k} \in \mathbb{R}$ is the kth diagonal element of $\mathbf{P}_m$. And $z_k \in \mathbb{C}$ represents the additive white Gaussian noise (AWGN) for the kth user, with $z_k \sim \mathcal{CN}(0, \sigma_{\mathrm{noise}}^2)$ with σ_{noise} being the standard noise deviation.

III. Energy Efficiency Under mmWave Heterogeneous Cell-Free Systems

In this section, we derive the EE expression for mmWave heterogeneous cell-free systems.

A. Instantaneous Energy Efficiency Formulation

In heterogeneous cell-free systems, the instantaneous sum rate and power consumption are achieved on the premise that the APs states are provided due to the fact that various types of APs are deployable. Let $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{H}}$ denote the set of types of APs and channel matrices, respectively, such that $\boldsymbol{\mathcal{S}} = \{\boldsymbol{\mathcal{S}}_1, \cdots, \boldsymbol{\mathcal{S}}_T\}$ and $\boldsymbol{\mathcal{H}} = \{\mathbf{H}_1, \cdots, \mathbf{H}_M\}$. Based on the received signal in (3), the achievable rate of the kth user can be expressed as

$$R_k\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right) = B\log_2\left(1 + A_k^* R_{\mathrm{IF}}^{-1} A_k\right), \tag{4}$$

where B is the bandwidth (Hz) and $A_k = \sum_{m\in\boldsymbol{\mathcal{A}}_k} \mathbf{H}_{m,k}^{\mathrm{T}}\mathbf{F}_m\mathbf{W}_{m,k}\sqrt{P_{m,k}}$. Moreover, the power of interference and noise is given as

$$R_{\mathrm{IF}} = \sum_{k'\neq k}\sum_{m'\in\mathcal{A}_{k'}} \left(\mathbf{H}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}\sqrt{P_{m',k'}}\right) \times \left(\mathbf{H}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}\sqrt{P_{m',k'}}\right)^* + \sigma_{\mathrm{noise}}^2. \tag{5}$$

The number of users is far less than the number of APs, and the receiver architectures for users are miniaturized, i.e., there is only minimal power consumption at the user equipment (UE) side, which can be neglected. Therefore, focusing on the active components of the transmitter, the power consumption of the mth AP is found to be

$$P_m^{\mathrm{C}}\left(\mathbf{H}_m|\boldsymbol{\mathcal{S}}_t\right) = P_{\mathrm{BB}} + l_t P_{\mathrm{RF}} + N_{\mathrm{PS},t}P_{\mathrm{PS}} + \sum_{n=1}^{N_t} P_{\mathrm{PA}}\left(P_n^{\mathrm{out}}\right), \tag{6}$$

where $N_{\mathrm{PS},t}$ is the number of phase shifters. This is a function of l_t and N_t, which is related to the transmitter architecture. For example, we have $N_{\mathrm{PS},t} = l_t N_t$ under the fully connected architecture, and $N_{\mathrm{PS},t} = N_t$ under the subarray architecture. Besides, P_{BB}, P_{RF} and P_{PS} are power costs of the baseband, an RF chain, and a phase shifter, respectively. The function $P_{\mathrm{PA}}(\cdot)$ represents the power consumption of a power amplifier (PA) at the nth transmit antenna, and it is given as the product of the output power (P_n^{out}) and the PA efficiency. Generally, the PA efficiency model can be divided into linear and nonlinear components. Furthermore, the linear PA efficiency often exists in the lower-frequency communication systems, while the nonlinear PA efficiency exists in the mmWave and sub-Terahertz bands.

Therefore, the instantaneous EE of the heterogeneous cell-free systems can be written as

$$\xi_{\mathrm{EE}}\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right) = \frac{R\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)} \left(\frac{\mathrm{bits}}{\mathrm{Joule}}\right), \tag{7}$$

where $R(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})$ is the sum achievable rate (bits/s) with $R(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}) = \sum_{k=1}^{K} R_k(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})$. Similarly, $P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})$ is the total power consumption (Watts) and $P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}) = \sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M} \epsilon(|\boldsymbol{\mathcal{A}}_m^{-1}| - 1)P_m^{\mathrm{C}}(\mathbf{H}_m|\boldsymbol{\mathcal{S}}_t)$. $\epsilon(\cdot)$ is the traditional step function.

As illustrated in Introduction, the instantaneous EE helps to design the instantaneous signal process schemes like hybrid precoder and power allocation, etc. However, considering that the heterogeneous cell-free systems are deployable, i.e., immobile only for a short period of time, it is necessary to measure the EE performance from the statistical aspect.

B. Ergodic Energy Efficiency Formulation

According to the analyses in Section III-A, the sum rate and power consumption are functions of deployable APs. Generally, the hybrid pre-coding and power allocation matrices are designed according to the channel matrix. Furthermore, the output power at each transmit antenna is decided by the hybrid pre-coding and power allocation matrices, which then affects the power consumption of a PA. So the sum rate and power consumption are also functions of channel matrices.

Based on the instantaneous EE in (7), the ergodic EE of the deployable heterogeneous cell-free systems is defined as

$$\bar{\xi}_{\mathrm{EE}}\left(\boldsymbol{\mathcal{S}}\right) = \mathrm{E}_{\boldsymbol{\mathcal{H}}}\left[\frac{R\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)}\right] \left(\frac{\mathrm{bits}}{\mathrm{Joule}}\right), \tag{8}$$

where $\mathrm{E}_{\boldsymbol{\mathcal{H}}}[\cdot]$ represents the expectation of " $\cdot$ " over $\boldsymbol{\mathcal{H}}$. Considering that the channel model consists of mutually independent large-scale fading and small-scale fading, let $\boldsymbol{\mathcal{L}} = \{\mathbf{L}_1, \cdots, \mathbf{L}_M\}$ denote the set of large-scale fading matrices, and $\langle = \{\mathbf{h}_1, \cdots, \mathbf{h}_M\}$ the set of small-scale fading matrices. Then, (8) can be transformed to

$$\bar{\xi}_{\mathrm{EE}}\left(\boldsymbol{\mathcal{S}}\right) = \mathrm{E}_{\boldsymbol{\mathcal{L}}}\left[\mathrm{E}_{\langle}\left[\frac{R\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)}\right]\right], \tag{9}$$

in which $\mathrm{E}_{\boldsymbol{\mathcal{L}}}[.]$ and $\mathrm{E}_{\langle}[.]$ are the expectation functions of $\boldsymbol{\mathcal{L}}$ and $\langle$, respectively.

Through integration, (9) becomes (10), where $f(\mathbf{L}_1, \cdots, \mathbf{L}_M)$ and $f(\mathbf{h}_1, \cdots, \mathbf{h}_M)$ are the joint probability density functions (PDFs) of $\{\mathbf{L}_1, \cdots, \mathbf{L}_M\}$ and $\{\mathbf{h}_1, \cdots, \mathbf{h}_M\}$, respectively.

As shown in Table I, one feasible way to achieve high EE consists of two steps. The normalized hybrid pre-coding matrices are designed to eliminate the interference between multiple users and data streams, and the optimal power allocation is calculated to maximize EE. The unitary process is followed when APs and users are both equipped with multiple antennas, and the zero-forcing (ZF) process is taken when each AP or user is equipped with a single antenna.

$$\bar{\xi}_{\mathrm{EE}}\left(\boldsymbol{\mathcal{S}}\right)=\int_{\mathbf{L}_M}\cdots\int_{\mathbf{L}_1}\left(\int_{\mathbf{h}_M}\cdots\int_{\mathbf{h}_1}\frac{R\left(\{\mathbf{h}_M\sqrt{\mathbf{L}_M},\cdots,\mathbf{h}_1\sqrt{\mathbf{L}_1}\}|\boldsymbol{\mathcal{S}}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\{\mathbf{h}_M\sqrt{\mathbf{L}_M},\cdots,\mathbf{h}_1\sqrt{\mathbf{L}_1}\}|\boldsymbol{\mathcal{S}}\right)}f\left(\mathbf{h}_1,\cdots,\mathbf{h}_M\right)d_{\mathbf{h}_1}\cdots d_{\mathbf{h}_M}\right)\times f\left(\mathbf{L}_1,\cdots,\mathbf{L}_M\right)d_{\mathbf{L}_1}\cdots d_{\mathbf{L}_M} \tag{10}$$

$$\bar{\xi}_{\mathrm{EE}}\left(\boldsymbol{\mathcal{S}}\right)=\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_m^{-1}}^{m=M}\int_{L_{m,k}}\right)\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_m^{-1}}^{m=M}\int_{\mathbf{h}_{m,k}}\right)\frac{B\sum_{k=1}^{K}\log_2\left(1+\gamma_k\right)}{P_{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}\right)+P_{\mathrm{V}}\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)}\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_m^{-1}}^{m=M}f\left(\mathbf{h}_{m,k}\right)\right)\times\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_m^{-1}}^{m=M}d_{\mathbf{h}_{m,k}}\right)\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_m^{-1}}^{m=M}f(L_{m,k})\right)\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_m^{-1}}^{m=M}d_{L_{m,k}}\right) \tag{16}$$

Compared with large-scale fading, small-scale fading plays a significant role in interference. In order to eliminate interference, the digital part of the ZF hybrid precoder at the mth AP is given as

$$\mathbf{W}_m=\widetilde{\mathbf{W}}_m\boldsymbol{\Lambda}_m=\left(\mathbf{h}_m^{\mathrm{T}}\mathbf{F}_m\right)^{\mathrm{H}}\left(\left(\mathbf{h}_m^{\mathrm{T}}\mathbf{F}_m\right)\left(\mathbf{h}_m^{\mathrm{T}}\mathbf{F}_m\right)^{\mathrm{H}}\right)^{-1}\boldsymbol{\Lambda}_m, \tag{11}$$

where $\boldsymbol{\Lambda}_m$ is a normalized matrix which satisfies $\|\mathbf{F}_m\mathbf{W}_{m,k}\|_{\mathrm{F}}^2=1$ with $k\in\mathcal{A}_m^{-1}$. $\boldsymbol{\Lambda}_m$, in turn, can be written as $\mathrm{diag}(\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,1}\|_{\mathrm{F}}^{-1},\cdots,\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,|\mathcal{A}_m^{-1}|}\|_{\mathrm{F}}^{-1})$. The analog part of the ZF hybrid precoder is often set to be the matched filter (MF), given as $\mathbf{F}_m=\mathrm{angle}(\mathbf{h}_m^*)$. This allows us to transform the sum rate as,

$$R\left(\{\mathbf{h}_M\sqrt{\mathbf{L}_M},\cdots,\mathbf{h}_1\sqrt{\mathbf{L}_1}\}|\boldsymbol{\mathcal{S}}\right)=B\sum_{k=1}^{K}\log_2\left(1+\gamma_k\right), \tag{12}$$

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

$$\gamma_k=\frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{P_{m,k}L_{m,k}}/\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\mathrm{F}}|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2} \tag{13}$$

The interference part P_k^{I} exists for the ZF hybrid precoder locally at each AP. As $m'\in\boldsymbol{\mathcal{S}}_{t'}$, we can see this interference as $P_k^{\mathrm{I}}=\sum_{m'\notin\boldsymbol{\mathcal{A}}_k}\sum_{k'\in\boldsymbol{\mathcal{A}}_{m'}^{-1}}|\sqrt{P_{m',k'}L_{m',k'}}\mathbf{h}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}|^2$.

From (6), we can see that the sum power consumption can be divided into the constant and the variable parts of the channel matrix. The former can be written as

$$P_{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}\right)=\sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\epsilon\left(|\boldsymbol{\mathcal{A}}_m^{-1}|-1\right)\left(P_{\mathrm{BB}}+l_tP_{\mathrm{RF}}+N_{\mathrm{PS},t}P_{\mathrm{PS}}\right), \tag{14}$$

while the latter can be written as

$$P_{\mathrm{V}}\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)=\sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\sum_{n=1}^{N_t}P_{\mathrm{PA}}\left(\|\left(\mathbf{F}_m\mathbf{W}_m\sqrt{\mathbf{P}_m}\right)_n^{\mathrm{T}}\|_{\mathrm{F}}^2\right), \tag{15}$$

where $(\mathbf{F}_m\mathbf{W}_m\sqrt{\mathbf{P}_m})_n^{\mathrm{T}}$ is the nth column of $(\mathbf{F}_m\mathbf{W}_m\sqrt{\mathbf{P}_m})^{\mathrm{T}}$ and $\|(\mathbf{F}_m\mathbf{W}_m\sqrt{\mathbf{P}_m})_n^{\mathrm{T}}\|_{\mathrm{F}}^2$ is the power output of the nth antenna at the mth AP, i.e., P_n^{out} in (6).

When the spatial conditions are satisfied (i.e., when APs and users are dispersedly distributed), we see that the fading is independent between the various APs and users. This means that the joint probability density functions $f(\mathbf{L}_1,\cdots,\mathbf{L}_M)$ and $f(\mathbf{h}_1,\cdots,\mathbf{h}_M)$ can now be written as $f(\mathbf{L}_1)\cdots f(\mathbf{L}_M)$ and $f(\mathbf{h}_1)\cdots f(\mathbf{h}_M)$, respectively. These can be further written as $f(\mathbf{L}_m)=f(L_{m,1})\cdots f(L_{m,|\boldsymbol{\mathcal{A}}_m^{-1}|})$ and, following which, $f(\mathbf{h}_m)=f(\mathbf{h}_{m,1})\cdots f(\mathbf{h}_{m,|\boldsymbol{\mathcal{A}}_m^{-1}|})$.

Therefore, (10) can be rewritten into (16), where $\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_m^{-1}}^{m=M}$ is the successive multiplication operation. Eqn. (16) gives the expression for the ergodic EE of the deployable heterogeneous cell-free systems with ZF hybrid precoders.

IV. Closed Form of Ergodic Energy Efficiency

In this section, a semi-closed expression of the ergodic EE is given first after several simplification methods. Based on it, two cases of heterogeneous gain are also derived in order to have direct insights into the ergodic EE. Further, we theoretically analyze the effects of the main simplification method on the ergodic EE performance.

A. Semi-Closed Expression of Ergodic EE

Due to the difficulty in directly obtaining instructive conclusions from (16), it is acceptable to make a few mathematical simplifications to the expression of the ergodic EE.

Considering a fully-connected architecture, here, we assume $N_{\mathrm{PS},t}=l_tN_t$. Since each user has a singular data stream, the number of working RF chains is equal to that of serving users. Hence, the constant part of the sum power consumption becomes

$$P_{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}\right)=\sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\epsilon\left(|\boldsymbol{\mathcal{A}}_m^{-1}|-1\right)\times\left(P_{\mathrm{BB}}+|\boldsymbol{\mathcal{A}}_m^{-1}|P_{\mathrm{RF}}+|\boldsymbol{\mathcal{A}}_m^{-1}|N_tP_{\mathrm{PS}}\right). \tag{17}$$

Following from the workings in [26], using the traditional system parameters, the nonlinear PA efficiency can be relaxed to a linear function within an acceptable error. Specially, the maximum total error caused by the linearization operation of PA efficiency is $\left(N_{\mathrm{A}}P_{\max}^{\mathrm{PA}}\right)/\left(8\pi\right)$, where $P_{\max}^{\mathrm{PA}}$ is the maximum output power of a PA and N_{A} is the number of total transmit antennas. With the typical parameters, i.e., $N_{\mathrm{A}}=128$ and $P_{\max}^{\mathrm{PA}}=100$ mW, power consumptions

of the baseband, RF chain, phase shifter are 200 mW, 120 mW, 20 mW, respectively, we can get $P_{\mathrm{C}}(\boldsymbol{\mathcal{S}})$ of (14) and $P_{\mathrm{V}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})$ of (15) in the manuscript. Then it is able to calculate that $(N_{\mathrm{A}}P_{\max}^{\mathrm{PA}})/(8\pi)$ accounts for a very small part of the total power consumption $(P_{\mathrm{C}}(\boldsymbol{\mathcal{S}})+P_{\mathrm{V}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}))$, which approximately equals to 3.52%. So the linearization of a PA efficiency is acceptable.

Assuming ρ_t to be the approximate PA efficiency coefficient taken by the APs in $\boldsymbol{\mathcal{S}}_t$, the variable part of the sum power consumption in (15) can be simplified to

$$P_{\mathrm{V}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}) \approx \sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M} \epsilon\left(|\boldsymbol{\mathcal{A}}_m^{-1}|-1\right)\left(\sum_{k\in\boldsymbol{\mathcal{A}}_m^{-1}} P_{m,k}\rho_t\right). \tag{18}$$

As the power allocation matrix is a joint function of the small-scale and large-scale fadings, it leads to difficulty in analyzing $\bar{\xi}_{\mathrm{EE}}(\boldsymbol{\mathcal{S}})$. Taking the power to be discrete, with $m\in\boldsymbol{\mathcal{S}}_t$ and $k\in\boldsymbol{\mathcal{A}}_m^{-1}$, $P_{m,k}$ falls in the range $[0,P_t]$, and $[0,P_t]$ is equispaced divided into I segments.

When $P_{m,k}$ is located in the ith segment, we set $P_{m,k}$ to be the sum of its small initial value $(P_{m,k}^i)$ and the error $(\Delta_{m,k}^i)$, i.e., $P_{m,k}=P_{m,k}^i+\Delta_{m,k}^i$, where $0\le\Delta_{m,k}^i\le P_t/I$. The error between $P_{m,k}$ and $P_{m,k}^i$ decreases with increasing of I. When I is large enough, with $(1+x)^{1/2}\approx 1+(1/2)x$, it is easy to deduce that $\sqrt{P_{m,k}}=\sqrt{P_{m,k}^i+\Delta_{m,k}^i}\approx\sqrt{P_{m,k}^i}+(1/2)\Delta_{m,k}^i/\sqrt{P_{m,k}^i}$. Next, let $\delta_{m,k}^i=1$ on the premise of $P_{m,k}=P_{m,k}^i+\Delta_{m,k}^i$; otherwise, $\delta_{m,k}^i$ is set as $\delta_{m,k}^i=0$. Hence, we can derive $\sqrt{P_{m,k}}\approx\sum_{i=1}^{I}\delta_{m,k}^i\sqrt{P_{m,k}^i}+\sum_{i=1}^{I}\delta_{m,k}^i(1/2)\Delta_{m,k}^i/\sqrt{P_{m,k}^i}$.

In a similar manner, we are able to arrive at that $\sqrt{L_{m,k}}\approx\sum_{o=1}^{O}\delta_{m,k}^o\sqrt{L_{m,k}^o}+\sum_{o=1}^{O}\delta_{m,k}^o(1/2)\Delta_{m,k}^o/\sqrt{L_{m,k}^o}$. Here, O is the number of segments dividing the value range of $L_{m,k}$. Set $\delta_{m,k}^o=1$ if $L_{m,k}=L_{m,k}^o+\Delta_{m,k}^o$; otherwise, let $\delta_{m,k}^o=0$. And here, $L_{m,k}^o$ is the left starting point of the oth segment and $\Delta_{m,k}^o$ is the corresponding error.

Similarly, we are able to derive the discrete $\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\mathrm{F}}^{-1}$ and $|\mathbf{h}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}|^2$. Denoting the number of segments by J and L, let $\eta_{m,k}^j$ (or $\mu_{m',k',k}^l$) be the left starting point of the j(or l)th segment and $\Delta_{m,k}^j$ (or $\Delta_{m',k',k}^l$) be the corresponding error. Set $\delta_{m,k}^j=1$ (or $\delta_{m',k',k}^l=1$) if $\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\mathrm{F}}^{-1}=\eta_{m,k}^j+\Delta_{m,k}^j$ (or $|\mathbf{h}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}|^2=\mu_{m',k',k}^l+\Delta_{m',k',k}^l$); otherwise, $\delta_{m,k}^j=0$ (or $\delta_{m',k',k}^l=0$). This gives us $\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\mathrm{F}}^{-1}=\sum_{j=1}^{J}\delta_{m,k}^j\eta_{m,k}^j+\sum_{j=1}^{J}\delta_{m,k}^j\Delta_{m,k}^j$ and $|\mathbf{h}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}|^2=\sum_{l=1}^{L}\delta_{m',k',k}^l\mu_{m',k',k}^l+\sum_{l=1}^{L}\delta_{m',k',k}^l\Delta_{m',k',k}^l$. Effectiveness of discretization will be analyzed in Section IV-C and verified by simulation results in Section V-A.

With large enough I, O, J, and L, we are able to derive an approximate expression of γ_k, that is

$$\gamma_k\approx\frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{P_{m,k}^{\mathrm{U}}}|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2} \tag{19}$$

in which $\sqrt{P_{m,k}^{\mathrm{U}}}$ is equal to $(\sum_{o=1}^{O}\delta_{m,k}^o\sqrt{L_{m,k}^o})(\sum_{i=1}^{I}\delta_{m,k}^i\sqrt{P_{m,k}^i})(\sum_{j=1}^{J}\delta_{m,k}^j\eta_{m,k}^j)$. Moreover, P_k^{I} is equal to $\sum_{m'\notin\boldsymbol{\mathcal{A}}_k}\sum_{k'\in\boldsymbol{\mathcal{A}}_{m'}^{-1}}(\sum_{i'=1}^{I}\delta_{m,k}^{i'}\sqrt{P_{m',k'}^{i'}})^2(\sum_{o'=1}^{O}\delta_{m',k'}^{o'}\sqrt{L_{m',k'}^{o'}})^2(\sum_{l=1}^{L}\delta_{m',k',k}^l\mu_{m',k',k}^l)$, which represents the interference. Hence, the approximate power consumption becomes

$$\begin{aligned}P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}) \approx \sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\epsilon\left(|\boldsymbol{\mathcal{A}}_m^{-1}|-1\right)\Big(P_{\mathrm{BB}}+\\ |\boldsymbol{\mathcal{A}}_m^{-1}|P_{\mathrm{RF}}+|\boldsymbol{\mathcal{A}}_m^{-1}|N_tP_{\mathrm{PS}}+\sum_{k\in\boldsymbol{\mathcal{A}}_m^{-1}}\sum_{i=1}^{I}\delta_{m,k}^iP_{m,k}^i\rho_t\Big).\end{aligned} \tag{20}$$

The probability of the case where $\delta_{m,k}^i=1$, $\delta_{m',k'}^{i'}=1$, $\delta_{m,k}^j=1$, $\delta_{m',k',k}^l=1$, $\delta_{m,k}^o=1$ and $\delta_{m',k'}^{o'}=1$ is denoted as $p(i,i',j,l,o,o')$. Then discrete ergodic EE can be therefore expressed as,

$$\bar{\xi}_{\mathrm{EE}}^{\mathrm{D}}(\boldsymbol{\mathcal{S}})=B\sum_{i=1}^{I}\sum_{i'=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{L}\sum_{o=1}^{O}\sum_{o'=1}^{O}p(i,i',j,l,o,o')\frac{\sum_{k=1}^{K}\log_2\left(1+\frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{L_{m,k}^oP_{m,k}^i}\eta_{m,k}^j|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2}\right)}{P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})} \tag{21}$$

where $P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})=\sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\epsilon(|\boldsymbol{\mathcal{A}}_m^{-1}|-1)\times(P_{\mathrm{BB}}+|\boldsymbol{\mathcal{A}}_m^{-1}|P_{\mathrm{RF}}+|\boldsymbol{\mathcal{A}}_m^{-1}|N_tP_{\mathrm{PS}}+\sum_{k\in\boldsymbol{\mathcal{A}}_m^{-1}}P_{m,k}^i\rho_t)$ and P_k^{I} is equal to $\sum_{m'\notin\boldsymbol{\mathcal{A}}_k}\sum_{k'\in\boldsymbol{\mathcal{A}}_{m'}^{-1}}P_{m',k'}^{i'}L_{m',k'}^{o'}\mu_{m',k',k}^l$.

We know that a logarithmic function can be approximated by a piecewise function, allowing us to express the sum rate by a piecewise linear function of SINR, i.e., $\log_2(1+\gamma_k)\approx\beta_s\gamma_k+C_s$. We have β_s $(1\le s\le S)$ as the slope of the line in the sth segment and C_s as the constant in the sth segment, with the total number of segments denoted as S. When the number of segments is large enough, a logarithmic function can be replaced by a piecewise function with a negligible error. The partition of segments can be uniform and non-uniform, which is decided upon by a design rule [39]. The non-uniform partition can focus on the value range of SINR with high probability at the cost of complexity, while the uniform partition is simple. Following this, (21) becomes

$$\bar{\xi}_{\mathrm{EE}}^{\mathrm{D}}(\boldsymbol{\mathcal{S}})\approx B\sum_{i=1}^{I}\sum_{i'=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{L}\sum_{o=1}^{O}\sum_{o'=1}^{O}p(i,i',j,l,o,o')\frac{\sum_{k=1}^{K}\sum_{s=1}^{S}\left(\beta_s\frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{L_{m,k}^oP_{m,k}^i}\eta_{m,k}^j|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2}+C_s\right)\delta_{k,s}}{P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})} \tag{22}$$

where $\delta_{k,s}=1$ when γ_k lies in the sth segment. Otherwise, $\delta_{k,s}=0$. This semi-closed expression is useful for analyzing the ergodic EE of heterogeneous cell-free systems. Based on it, heterogeneous gain will be discussed in Section IV-B. And several insights will be given in Sections V-B, V-C, and V-D, where a part of easily configured variables are taken as examples, i.e., the transmit power, and the numbers of APs and users, which can guide the selection of system parameters.

The complexity of discretization of the integral is measured by the time complexity, which is related to the number of cyclic operation. According to the semi-closed ergodic EE in (22), calculation of both the instantaneous EE and the probability of discretized variables are required to realize discretization of the integral. And the probability of discretized variables can be given through realizing instantaneous EE of discretized variables for many times. Then it is necessary to calculate the number of cyclic operation required to achieve instantaneous EE of discretized variables. Since each AP independently designs the hybrid pre-coding and power allocation matrix, which are supposed to be achieved by the closed formulas, it can be considered to circulate M times. Then the corresponding time complexity can be written as $\mathcal{O}(NM)$, where N is the number of channel realizations. The method of bisection is used to find the order number of segment where a discretized variable lies. Then the maximum time complexity of discretization of variables is expressed as $\mathcal{O}(N\log_2 O+N\log_2 J+N\log_2 L+N\log_2 I+N\log_2 S)$. Therefore, the whole time complexity to achieve instantaneous EE of discretized variables is $\mathcal{O}(N\log_2 O+N\log_2 J+N\log_2 L+N\log_2 I+N\log_2 S+NM)$, which is also the whole time complexity of discretization of the integral. Generally, the number of channel realizations (N) plays a main role in the whole time complexity. As shown in the simulation results of Figs. 2 and 3, when the number of channel realizations exceeds a threshold, such as 100, the fluctuation caused by discretization tends to be stable. Therefore, the complexity of the discretization of the integral for evaluating the ergodic EE is acceptable.

B. Heterogeneous Gain

Generally, the AP with sophisticated architecture performs powerful capability, and the parameters, such as P_t, N_t, l_t and etc, are often larger than those of APs with simpler architecture. Thus, this tends to have an increased manufacturing complexity and cost, leading to fewer number of APs in heterogeneous cell-free systems. Therefore, the ergodic EEs caused by the APs with the most sophisticated architectures and the APs with the simplest architectures represent two extremes of ergodic EEs in heterogeneous cell-free systems. The larger the difference between the two extremes, the more ergodic EE gain is achieved by the heterogeneous cell-free systems. A heterogeneous gain is used to describe the difference between two extremes. Since the APs with the most sophisticated architectures own the maximum parameters, such as P_t. Let $\boldsymbol{\mathcal{S}}_{\max}=\boldsymbol{\mathcal{S}}_{t^*}$ and $\boldsymbol{\mathcal{S}}_{\min}=\boldsymbol{\mathcal{S}}_{t'}$, where $t^*=\max_t\{P_1,\cdots,P_t,\cdots,P_T\}$ and $t'=\min_t\{P_1,\cdots,P_t,\cdots,P_T\}$, respectively. In this manner, we have $\max\{a,b\}=a$ ($\min\{a,b\}=b$) when $a\geq b$; otherwise, we will have $\max\{a,b\}=b$ ($\min\{a,b\}=a$). Here, we define a heterogeneous gain as

$$\xi^{\mathrm{G}}=\frac{\bar{\xi}_{\mathrm{EE}}\left(\boldsymbol{\mathcal{S}}_{\max}\right)}{\bar{\xi}_{\mathrm{EE}}\left(\boldsymbol{\mathcal{S}}_{\min}\right)} \tag{23}$$

which indicates whether a heterogeneous system is necessary with respect to ergodic EE. If $\xi^{\mathrm{G}}>1$, it means that a heterogeneous system yields ergodic EE gain; otherwise, an isomorphic system does.

In fact, it is intractable to calculate the accurate ξ^{G}. From an engineering perspective, a simple and feasible solution to calculate ξ^{G} is desired. So, it is reasonable to calculate ξ^{G} using (22), called $\xi^{\mathrm{G,D}}$. This still proves to be quite difficult, and several simplification methods are described below.

In the low SINR region, it is easy to derive that $\log_2(1+\gamma_k)\approx\gamma_k$, which means $\beta_s=1$ and $C_s=0$. Here, the interference term can be ignored. Since β_s and C_s are both given, and the change in system parameters and signal processing mainly affects the useful signals, we chose to study the heterogeneous gain in the low SINR region. Now, (22) becomes

$$\begin{aligned}\bar{\xi}_{\mathrm{EE}}^{\mathrm{D}}\left(\boldsymbol{\mathcal{S}}\right)\approx & B\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{o=1}^{O}p\left(i,j,o\right)\\ & \frac{\sum_{k=1}^{K}|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{L_{m,k}^{o}P_{m,k}^{i}}\eta_{m,k}^{j}|^2}{\sigma_{\mathrm{noise}}^2P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}\right)}\end{aligned} \tag{24}$$

Since the power allocation is used to maximize the ergodic EE, equal power allocation is a lower bound. When equal power allocation is employed, the ergodic EE is further simplified to

$$\begin{aligned}\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}\left(\boldsymbol{\mathcal{S}}\right)= & \left(B/\sigma_{\mathrm{noise}}^2\right)\sum_{j=1}^{J}\sum_{o=1}^{O}p\left(j,o\right)\\ & \frac{\sum_{k=1}^{K}|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{(P_t/|\boldsymbol{\mathcal{A}}_m^{-1}|)L_{m,k}^{o}}\eta_{m,k}^{j}|^2}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}\right)}\end{aligned} \tag{25}$$

where $P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{S}})$ is written as $\sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\epsilon(|\boldsymbol{\mathcal{A}}_m^{-1}|-1)\times(P_{\mathrm{BB}}+|\boldsymbol{\mathcal{A}}_m^{-1}|P_{\mathrm{RF}}+|\boldsymbol{\mathcal{A}}_m^{-1}|N_tP_{\mathrm{PS}}+P_t\rho_t)$. Then it is reasonable to calculate ξ^{G} using (25), denoted as $\xi_{\mathrm{LW}}^{\mathrm{G,D}}=\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}}_{\max})/\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}}_{\min})$.

Since only a few APs own large and sophisticated architectures, longer distances between APs and users have a larger probability. That is, in comparison to $\boldsymbol{\mathcal{S}}_{\max}$, the advantage of using $\boldsymbol{\mathcal{S}}_{\min}$ lies in the smaller value of large-scale fading, experienced by the transmitted signals from APs belonging to $\boldsymbol{\mathcal{S}}_{\min}$. If the large-scale fadings are set to be the same for $\boldsymbol{\mathcal{S}}_{\max}$ and $\boldsymbol{\mathcal{S}}_{\min}$, this situation is beneficial for $\boldsymbol{\mathcal{S}}_{\max}$. Compared to small-scale fadings, large-scale fadings perform larger orders of magnitudes, which subsequently affects the received power. Thereafter, if the large-scale and small-scale fadings are set to be the same for $\boldsymbol{\mathcal{S}}_{\max}$ and $\boldsymbol{\mathcal{S}}_{\min}$, this situation is still beneficial for $\boldsymbol{\mathcal{S}}_{\max}$.

Therefore, by neglecting $L_{m,k}^{o}$ and $\eta_{m,k}^{j}$, we can have the first case of $\xi_{\mathrm{LW}}^{\mathrm{G,D}}$, that is

$$\begin{aligned}\xi_{\mathrm{LW}}^{\mathrm{G,D,1}}= & \frac{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}_{\min}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}_{\max}\right)}\times\\ & \frac{\sum_{k=1}^{K}|\sum_{m\in\boldsymbol{\mathcal{A}}_k,m\in\boldsymbol{\mathcal{S}}_{\max}}\sqrt{(P_{\max}/|\boldsymbol{\mathcal{A}}_m^{-1}|)}|^2}{\sum_{k'=1}^{K}|\sum_{m'\in\boldsymbol{\mathcal{A}}_{k'},m'\in\boldsymbol{\mathcal{S}}_{\min}}\sqrt{(P_{\min}/|\boldsymbol{\mathcal{A}}_{m'}^{-1}|)}|^2}\end{aligned} \tag{26}$$

where $P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{S}}_{\max})=\sum_{m=1}^{|\boldsymbol{\mathcal{S}}_{\max}|}\epsilon(|\boldsymbol{\mathcal{A}}_m^{-1}|-1)\times(P_{\mathrm{BB}}+|\boldsymbol{\mathcal{A}}_m^{-1}|P_{\mathrm{RF}}+|\boldsymbol{\mathcal{A}}_m^{-1}|N_{\max}P_{\mathrm{PS}}+P_{\max}\rho_{\max})$. $P_{\max}$, $N_{\max}$, and

$\rho_{\max}$ are the maximum output power, number of transmit antennas, and the PA efficiency coefficient corresponding to $\boldsymbol{\mathcal{S}}_{\max}$, respectively. Similarly, it is easy to derive $P_{\text{all}}^{\text{C}}(\boldsymbol{\mathcal{S}}_{\min})$.

Without the complicated CSI and statistical information, this version of the heterogeneous gain can help to determine whether a heterogeneous system is beneficial. The details are as follows,

Remark 1 (for the first case of the gain): When $\xi_{\text{LW}}^{\text{G,D},1} \approx 1$, it is better to employ an isomorphic system, whereas for $\xi_{\text{LW}}^{\text{G,D},1} << 1$ or $\xi_{\text{LW}}^{\text{G,D},1} >> 1$, it is reasonable to employ a heterogeneous system to achieve higher ergodic EE.

It is noted that $\xi_{\text{LW}}^{\text{G,D},1}$ is useful but not accurate enough for determining the necessity of a heterogeneous system. Since the large-scale and small-scale fadings are mutually independent, we have $p(j,o) = p(j)p(o)$. Recalling the definition of probability, $p(j) = N_j/N_J$ and $p(o) = N_o/N_O$, where N_j enumerates $\delta_{m,k}^j = 1$ and N_J enumerates small-scale fading realizations. Similarly, N_o enumerates $\delta_{m,k}^o = 1$ and N_O enumerates large-scale fading realizations. Now, (25) can be converted into

$$\bar{\xi}_{\text{EE}}^{\text{D,LW}}(\boldsymbol{\mathcal{S}}) = \left(\frac{B}{\sigma_{\text{noise}}^2}\right)\sum_{j=1}^{J}\sum_{o=1}^{O}\left(\frac{N_j}{N_J}\right)\left(\frac{N_o}{N_O}\right) \frac{\sum_{k=1}^{K}|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{(P_t/|\boldsymbol{\mathcal{A}}_m^{-1}|)L_{m,k}^o}\eta_{m,k}^j|^2}{P_{\text{all}}^{\text{C}}(\boldsymbol{\mathcal{S}})} \tag{27}$$

Based on $\bar{\xi}_{\text{EE}}^{\text{D,LW}}(\mathcal{S})$ in (27), we can obtain the second case of the heterogenous gain ξ^{G} as,

$$\xi_{\text{LW}}^{\text{G,D},2} = \frac{\bar{\xi}_{\text{EE}}^{\text{D,LW}}(\boldsymbol{\mathcal{S}}_{\max})}{\bar{\xi}_{\text{EE}}^{\text{D,LW}}(\boldsymbol{\mathcal{S}}_{\min})} \tag{28}$$

When N_J and N_O are set appropriately, the value of $\xi_{\text{LW}}^{\text{G,D},2}$ approaches $\xi_{\text{LW}}^{\text{G,D}}$, thus we can get a relatively accurate value of ξ^{G}. Therefore, $\xi_{\text{LW}}^{\text{G,D},2}$ is a stronger determinant for the necessity of a heterogeneous system.

Moreover, with the help of $\xi_{\text{LW}}^{\text{G,D},2}$, we present a heuristic method to guide the deployment of heterogeneous cell-free systems: Divide the users into T clusters according to a clustering algorithm. Hence, the whole coverage area is divided into corresponding T parts. In each part of the area needed to be deployed, successively change $\boldsymbol{\mathcal{S}}_{\max} = \boldsymbol{\mathcal{S}}_t$ according to $\boldsymbol{\mathcal{S}}$ while keeping $\boldsymbol{\mathcal{S}}_{\min}$ constant, and further calculate the corresponding $\xi_{\text{LW}}^{\text{G,D},2}$. The $\boldsymbol{\mathcal{S}}_{\max}$ with the maximum $\xi_{\text{LW}}^{\text{G,D},2}$ is considered the recommended deployment in the corresponding part of the area. Thereafter, update $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{S}}/\boldsymbol{\mathcal{S}}_{\max}$ and $\boldsymbol{\mathcal{S}}_{\min}$ (one set in $\boldsymbol{\mathcal{S}}$ with the minimum transmit power). The remaining parts of the area are also to be updated by removing the already deployed parts. Repeat the forward steps until each part of the area owns a serviced AP set. APs in a serviced AP set are uniformly distributed in the corresponding part of the area.

Remark 2 (for the second case of the gain): If $\xi_{\text{LW}}^{\text{G,D},2} > 1$, it is necessary to employ a heterogeneous system to achieve higher ergodic EE. Larger the $\xi_{\text{LW}}^{\text{G,D},2}$, the better EE performance. Besides, $\xi_{\text{LW}}^{\text{G,D},2}$ is useful for guiding the deployment of heterogeneous cell-free systems.

Although **Remark 1** and **Remark 2** were derived in the low SINR regions, they remain valid in the high SINR regions, as verified by the simulation results.

C. Effect of Discretization

It is necessary to analyze the effects of the simplification methods on the system performance. As shown in Section III-A, the main simplification method was to discretize the variates and statistical information in the ergodic EE. Since the heterogeneous gain indicates the ergodic EE of the heterogeneous cell-free systems, we focus on the effect of discretization on it.

Based on Eqns. (21) and (23), the heterogeneous gain with discretization can be further expressed as

$$\begin{aligned}\xi^{\text{G,D}} &= \frac{\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\max})}{\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\min})} = \frac{\bar{\xi}_{\text{EE}}(\boldsymbol{\mathcal{S}}_{\max}) - \Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\max})}{\bar{\xi}_{\text{EE}}(\boldsymbol{\mathcal{S}}_{\min}) - \Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\min})} \\ &= \xi^{\text{G}}\frac{1 - \Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\max})/\bar{\xi}_{\text{EE}}(\boldsymbol{\mathcal{S}}_{\max})}{1 - \Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\min})/\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\min})}\end{aligned} \tag{29}$$

where $\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\max}) = \bar{\xi}_{\text{EE}}(\boldsymbol{\mathcal{S}}_{\max}) - \Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\max})$. Here, $\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\max})$ is the ergodic EE with discretization; $\bar{\xi}_{\text{EE}}(\boldsymbol{\mathcal{S}}_{\max})$ is the ergodic EE; $\Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\max})$ is the discretization error. In a similar manner, this can be derived using $\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\min})$, $\bar{\xi}_{\text{EE}}(\boldsymbol{\mathcal{S}}_{\min})$, and $\Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\min})$. Especially, for the special case when $\Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\max}) = 0$ and $\Delta\bar{\xi}_{\text{EE}}^{\text{D}}(\boldsymbol{\mathcal{S}}_{\min}) = 0$, we can have $\xi^{\text{G,D}} = \xi^{\text{G}}$.

To have a direct insight into the effect of discretization on the heterogeneous gain, here, we use (27) to calculate $\xi^{\text{G,D}}$. Since the changes in the small-scale fading are faster than that in the large-scale fading, it is reasonable to only consider the discretization of variates influenced by the former. In other words, the large-scale fading is assumed to be constant. Then we are able to arrive at

$$\bar{\xi}_{\text{EE}}^{\text{D,LW}}(\boldsymbol{\mathcal{S}}|L_{m,k}^o, \forall k,m) = \left(\frac{B}{\sigma_{\text{noise}}^2}\right)\sum_{j=1}^{J}\left(\frac{N_j}{N_J}\right) \frac{\sum_{k=1}^{K}|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{(P_t/|\boldsymbol{\mathcal{A}}_m^{-1}|)L_{m,k}^o}\eta_{m,k}^j|^2}{P_{\text{all}}^{\text{C}}(\boldsymbol{\mathcal{S}})} \tag{30}$$

Recalling that $\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\text{F}}^{-1} = \eta_{m,k}$, denote its PDF by $f(\eta_{m,k})$. The ergodic EE with statistical information can then be written as (31).

The error in $\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\text{F}}^{-1} = \eta_{m,k}^j + \Delta_{m,k}^j$ caused by discretization, is expressed as (32). Since $\eta_{m,k}^j$ is given, $\Delta_{m,k}^j$ follows the distribution of $\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\text{F}}^{-1}$. According to the law of large numbers, it is reasonable to assume that $\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\text{F}}^{-1}$ follows a Gaussian random distribution, whose mean is $\mu_{m,k}$ and variance is $\sigma_{m,k}^2$. Hence, we have $\Delta_{m,k}^j \sim \mathcal{N}(\mu_{m,k} - \eta_{m,k}^j, \sigma_{m,k}^2)$.

For the sake of simplicity, we assume that each user is served by all the APs and each AP set only owns one element, which means $|\boldsymbol{\mathcal{A}}_m^{-1}| = K$ and $|\boldsymbol{\mathcal{S}}_t| = 1$. Additionally, $\mu_{m,k}$, $\sigma_{m,k}^2$, $\eta_{m,k}^j$, and $L_{m,k}^o$ are set to have the same for every user,

$$\bar{\xi}_{\mathrm{EE}}^{\mathrm{LW}}\left(\boldsymbol{\mathcal{S}}|L_{m,k}^{o},\forall k,m\right)=\left(\frac{B}{\sigma_{\mathrm{noise}}^{2}}\right)\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_{m}^{-1}}^{m=M}\int_{\eta_{m,k}}\right)\frac{\sum_{k=1}^{K}|\sum_{m\in\boldsymbol{\mathcal{A}}_{k}}\sqrt{(P_{t}/|\boldsymbol{\mathcal{A}}_{m}^{-1}|)L_{m,k}^{o}}\eta_{m,k}|^{2}}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}\right)}\times\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_{m}^{-1}}^{m=M}f\left(\eta_{m,k}\right)\right)\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_{m}^{-1}}^{m=M}d_{\eta_{m,k}}\right)\tag{31}$$

$$\Delta\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}\left(\boldsymbol{\mathcal{S}}|L_{m,k}^{o},\forall k,m\right)=\left(\frac{B}{\sigma_{\mathrm{noise}}^{2}}\right)\sum_{j=1}^{J}\left(\frac{N_{j}}{N_{J}}\right)\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_{m}^{-1}}^{m=M}\int_{\eta_{m,k}}\right)\frac{\sum_{k=1}^{K}|\sum_{m\in\boldsymbol{\mathcal{A}}_{k}}\sqrt{(P_{t}/|\boldsymbol{\mathcal{A}}_{m}^{-1}|)L_{m,k}^{o}}\Delta_{m,k}^{j}|^{2}}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}\right)}+\frac{\sum_{k=1}^{K}2\left(\sum_{m\in\boldsymbol{\mathcal{A}}_{k}}\sqrt{(P_{t}/|\boldsymbol{\mathcal{A}}_{m}^{-1}|)L_{m,k}^{o}}\eta_{m,k}^{j}\right)\left(\sum_{m\in\boldsymbol{\mathcal{A}}_{k}}\sqrt{(P_{t}/|\boldsymbol{\mathcal{A}}_{m}^{-1}|)L_{m,k}^{o}}\Delta_{m,k}^{j}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}\right)}\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_{m}^{-1}}^{m=M}f\left(\eta_{m,k}\right)\right)\left(\prod_{m=1,k\in\boldsymbol{\mathcal{A}}_{m}^{-1}}^{m=M}d_{\eta_{m,k}}\right)\tag{32}$$

and can now be denoted as μ_m, σ_m^2, η_m^j, and L_m^o, respectively. Then (31) becomes

$$\begin{aligned}\bar{\xi}_{\mathrm{EE}}^{\mathrm{LW}}\left(\boldsymbol{\mathcal{S}}_{t}|L_{t}^{o}\right)&=\left(\frac{B}{\sigma_{\mathrm{noise}}^{2}}\right)\frac{\sum_{k=1}^{K}\left(P_{t}/K\right)L_{t}^{o}\left(\mu_{t}^{2}+\sigma_{t}^{2}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}_{t}\right)}\\&=\left(\frac{B}{\sigma_{\mathrm{noise}}^{2}}\right)\frac{P_{t}L_{t}^{o}\left(\mu_{t}^{2}+\sigma_{t}^{2}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}_{t}\right)}\end{aligned}\tag{33}$$

Repeating this process, we obtain

$$\Delta\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}\left(\boldsymbol{\mathcal{S}}_{t}|L_{t}^{o}\right)=\left(\frac{B}{\sigma_{\mathrm{noise}}^{2}}\right)\frac{P_{t}L_{t}^{o}\left(\mu_{t}^{2}+\sigma_{t}^{2}-\sum_{j=1}^{J}\left(\frac{N_{j}}{N_{J}}\right)\left(\eta_{t}^{j}\right)^{2}\right)}{P_{\mathrm{all}}^{\mathrm{C}}\left(\boldsymbol{\mathcal{S}}_{t}\right)}\tag{34}$$

Further, (29) can be expressed as

$$\xi^{\mathrm{G,D}}=\xi^{\mathrm{G}}\frac{\left(\sum_{j=1}^{J}\left(\frac{N_{j}}{N_{J}}\right)\left(\eta_{\max}^{j}\right)^{2}\right)/\left(\mu_{\max}^{2}+\sigma_{\max}^{2}\right)}{\left(\sum_{j=1}^{J}\left(\frac{N_{j}}{N_{J}}\right)\left(\eta_{\min}^{j}\right)^{2}\right)/\left(\mu_{\min}^{2}+\sigma_{\min}^{2}\right)}\tag{35}$$

Although $\xi^{\mathrm{G,D}}$ can be larger than ξ^{G}, (35) does not indicate that the discretization method yields a gain. Factually, only the equality relationship between $\xi^{\mathrm{G,D}}$ and ξ^{G} is considered in our analysis. In particular, when J and N_J approach infinity, we can arrive that $\sum_{j=1}^{J}(\frac{N_j}{N_J})(\eta_{\max}^j)^2$ and $\sum_{j=1}^{J}(\frac{N_j}{N_J})(\eta_{\min}^j)^2$ equal to $\mu_{\max}^2+\sigma_{\max}^2$ and $\mu_{\min}^2+\sigma_{\min}^2$, respectively. Thus, we can see $\xi^{\mathrm{G,D}}=\xi^{\mathrm{G}}$. In other words, the discretization method makes sense only if both J and N_J are large enough.

V. Simulation Results and Discussions

In this section, we study the ergodic EE performance of the deployable heterogeneous cell-free systems. Consider a square area with a side length of 10000 meters. The channel model is set according to [38]. The approximate PA efficiency coefficient is given as $\rho_t = 4/\pi$ $(\forall t)$. The simulation parameters are detailed in Table II.

Furthermore, the numbers of users and types of APs are set as $K = 16$ and $T = 4$, respectively. For fairness, the relationships of parameters between various types of APs are $|\boldsymbol{\mathcal{S}}_{t-1}| = 2|\boldsymbol{\mathcal{S}}_t|$, $N_{t-1} = (1/2)N_t$, and $P_{t-1} = (1/2)P_t$. The ZF hybrid precoders with $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1$ and $\boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_3$ correspond to isomorphic systems, while those with $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_2, \boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$ and $\boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$ correspond to heterogeneous systems. Each user is served by all the APs, and each AP employs equal power allocation. A heterogeneous system is then deployed according to the proposed heuristic method in Section IV-B.

Simulation results in Fig. 2 are achieved by ξ^{G} and $\xi^{\mathrm{G,D}}$, where ξ^{G} is achieved by (23) with $\bar{\xi}_{\mathrm{EE}}(\boldsymbol{\mathcal{S}})$ showed in (16), and $\xi^{\mathrm{G,D}}$ is achieved by (23) with $\bar{\xi}_{\mathrm{EE}}^{\mathrm{D}}(\boldsymbol{\mathcal{S}})$ showed in (22). Fig. 3 is achieved by the discrete ergodic EE $\bar{\xi}_{\mathrm{EE}}^{\mathrm{D}}$ showed in (22) with various discrete steps. Average EE performances in Fig. 3-Fig. 12 are achieved by averaging the instantaneous EE shown in (7), with the number of channel realization being 1000. With given $\boldsymbol{\mathcal{S}}_{\max}$ and $\boldsymbol{\mathcal{S}}_{\min}$, the value of $\xi_{\mathrm{LW}}^{\mathrm{G,D},1}$ is achieved according to (26) and value of $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ is achieved according to (28). Then Fig. 13 is obtained by using $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$.

TABLE II
Simulation parameters

Parameters	Values
Power consumption of baseband	200 mW
Power consumption of an RF chain	120 mW
Power consumption of a phase shifter	20 mW
Bandwidth	100 MHz
Carrier frequency	30 GHz

A. Feasibility of Discretization

The feasibility of discretization is analyzed in Fig. 2 to investigate heterogeneous gain. Since the value range of a variable is given, the effect of the discrete step on the heterogeneous gain is the same as that of the number of segments. As shown in Fig. 2, ratio between $\xi^{\mathrm{G,D}}$ and ξ^{G} fluctuates around one. When the discrete step equals 0.1 or 0.01, the ratio between $\xi^{\mathrm{G,D}}$ and ξ^{G} approaches one with the increase of the number of channel realizations, indicating that $\xi^{\mathrm{G,D}}$ can approach ξ^{G}. It keeps in line with the analysis in IV-C.

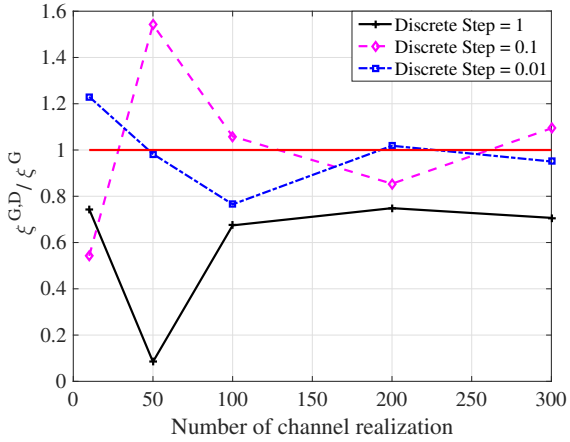


Fig. 2. Ratio between $\xi^{\mathrm{G,D}}$ and ξ^{G} with $\mathcal{S}_{\min} = \boldsymbol{\mathcal{S}}_1$ and $\mathcal{S}_{\max} = \boldsymbol{\mathcal{S}}_4$. $|\boldsymbol{\mathcal{S}}_1| = 32$, $N_1 = 4$, and $P_1 = -10$ dBm.

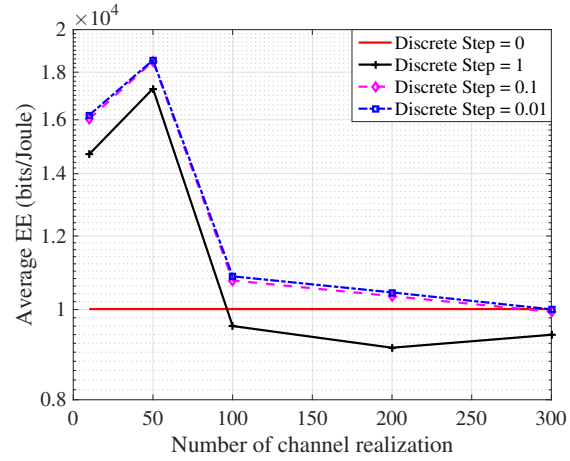


Fig. 3. Comparison between the semi-closed expression and the average instantaneous EE performance with various channel realizations, where $\mathcal{S}_{\min} = \boldsymbol{\mathcal{S}}_1$, $\mathcal{S}_{\max} = \boldsymbol{\mathcal{S}}_4$, $|\boldsymbol{\mathcal{S}}_1| = 32$, $N_1 = 4$, and $P_1 = -10$ dBm.

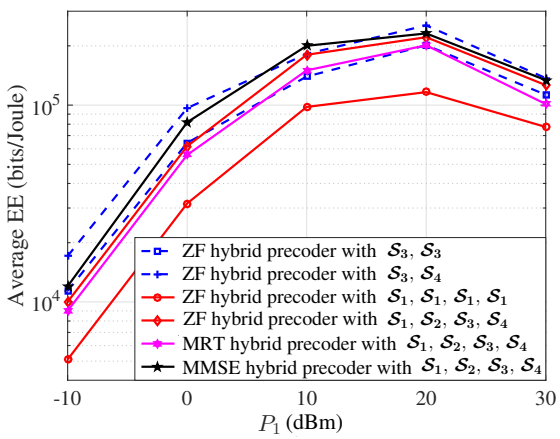


Fig. 4. Average EE performance with various transmit powers. Here, $|\boldsymbol{\mathcal{S}}_1| = 32$, and $N_1 = 4$.

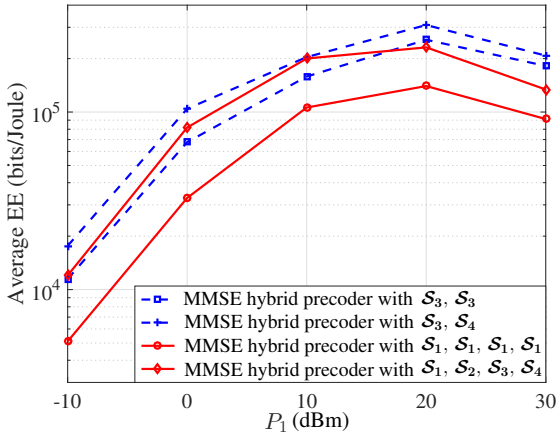


Fig. 5. Average EE performance with various transmit powers under MMSE based hybrid pre-coding. Here, $|\boldsymbol{\mathcal{S}}_1| = 32$, and $N_1 = 4$.

However, when the discrete step equals to 1, the ratio between $\xi^{\mathrm{G,D}}$ and ξ^{G} is always smaller than one, no matter what value of the number of channel realization is. The fundamental reason is that the loss on the sum rate caused by the large discrete step cannot be neglected. The similar phenomenon can be observed in Fig. 3, which compares values of the semi-closed expression from (22) and average EE (denoted as 'Discrete Step = 0'). With the enough small discrete step, when the number of channel realization increases, the value of semi-closed ergodic EE expression approaches that of the average EE, which verifies the effectiveness of the semi-closed expression. Besides, when the number of channel realizations exceeds the threshold, such as 100, the fluctuation tends to be stable. Therefore, we select representative $\{0.1, 50\}$, $\{0.01, 50\}$, and $\{0.01, 100\}$ as the discrete steps and realization numbers in the following simulations.

B. Effect of Transmit Power

Based on the semi-closed form approximation, it is able to obtain insight on the transmit power. When the transmit power is small, SINR is mainly affected by the useful signal power and linearly increases with the transmit power. The whole power consumption equals to the sum of the variable part and the constant part. The variable part is a linear function of the transmit power. Since the transmit power is small, the whole power consumption approximately equals to the constant part. So the ergodic EE performance increases in the low transmit power region. However, when the transmit power is large, SINR cannot linearly increase with the increasing of transmit power, but whole power consumption linearly increases with the increasing of transmit power. Then the ergodic EE decreases in the high transmit power region. Therefore, based on the semi-closed form, it is able to get that the ergodic EE performance first increases and then decreases with the increasing of transmit power, which is verified by Fig. 4.

Furthermore, as shown in Fig. 4, the average EE of the heterogeneous systems is higher than that of isomorphic systems. Besides, the performance gap between the ZF hybrid precoders with $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1$ and $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_2, \boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$ is larger than that between those with $\boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_3$ and $\boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$. For

the former group, i.e., $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1$ and $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_2, \boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$, $\boldsymbol{\mathcal{S}}_{\max} = \boldsymbol{\mathcal{S}}_4$ and $\boldsymbol{\mathcal{S}}_{\min} = \boldsymbol{\mathcal{S}}_1$, we calculated $\xi_{\mathrm{LW}}^{\mathrm{G,D},1}$ using (26) to be $[0.2311, 0.2311, 0.2305, 0.2253]$ with various transmit powers. For the latter group, i.e., $\boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_3$ and $\boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$, $\boldsymbol{\mathcal{S}}_{\max} = \boldsymbol{\mathcal{S}}_4$ and $\boldsymbol{\mathcal{S}}_{\min} = \boldsymbol{\mathcal{S}}_3$, then $\xi_{\mathrm{LW}}^{\mathrm{G,D},1}$ is calculated to $\xi_{\mathrm{LW}}^{\mathrm{G,D},1} = [0.5606, 0.5606, 0.5603, 0.5573]$. Hence, we found $\xi_{\mathrm{LW}}^{\mathrm{G,D},1}$ of the latter group to be larger than that of the former group. The average EE gap in the latter group was smaller, consistent with **Remark 1**.

However, in both groups, we have $\xi_{\mathrm{LW}}^{\mathrm{G,D},1} < 1$, which does not accurately reflect the necessity of a heterogeneous system. By setting the discrete step and realization number, we can compute $\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}})$ and hence $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$. Here, the discrete steps and realization numbers are set as $\{0.1, 50\}$, $\{0.01, 50\}$, and $\{0.01, 100\}$, respectively. For the former group, we are able to arrive at $\xi_{\mathrm{LW}}^{\mathrm{G,D},2} = [0.0297, 0.0552, 0.3559, 1.8078]$, $[0.3077, 5.0547, 2.3910, 3.1804]$, and $[6.8397, 4.4476, 3.5315, 4.9308]$. For the latter group, we have $\xi_{\mathrm{LW}}^{\mathrm{G,D},2} = [0.3130, 0.4547, 1.2171, 1.2602]$, $[1.4169, 1.7331, 1.1226, 1.7415]$, and $[1.1002, 1.5329, 1.2833, 1.1536]$. When the discrete step and realization number are set as $\{0.1, 50\}$, $\{0.01, 50\}$, the $\xi_{\mathrm{LW}}^{\mathrm{G,D},2} < 1$ case occurs, indicating the importance of setting an appropriate discrete step and realization number. When the discrete step and realization number are set as $\{0.01, 100\}$, we have the $\xi_{\mathrm{LW}}^{\mathrm{G,D},2} > 1$ case, implying that a heterogeneous system outperforms an isomorphic system, consistent with **Remark 2**. And $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ of the former group is larger than that of the latter group, the heterogeneous gain is bigger in the former group, as shown in simulation results. In the following, the discrete step and realization number are set as $\{0.01, 100\}$.

Considering the heterogeneous cell-free system with $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_2, \boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$, it is able to get average EEs with various hybrid pre-codings, such as ZF, maximum ratio transmission (MRT), minimum mean square error (MMSE) based hybrid pre-coding schemes. As show in Fig. 4, the MMSE hybrid pre-coding performs the best, then the ZF hybrid pre-coding, and finally the MRT hybrid pre-coding. Moreover, according to Appendix A, based on the semi-closed expressions of ergodic EEs for the MMSE hybrid pre-coding, the second version of heterogeneous gain is calculated for two groups, i.e., $\xi_{\mathrm{LW}}^{\mathrm{G,D},2} = \bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}}_4)/\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}}_1)$ and $\xi_{\mathrm{LW}}^{\mathrm{G,D},2} = \bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}}_4)/\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}}_3)$, which are $[4.9126, 3.9939, 2.5362, 3.0455]$, and $[2.1158, 1.9204, 1.6926, 1.5813]$, respectively. This indicates the gain brought by the first group $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_1$, $\boldsymbol{\mathcal{S}}_1$ and $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_2, \boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$ is larger than the gain brought by the second group $\boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_3$ and $\boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$, as shown in Fig. 5. By the same way, it is able to get the second version of heterogeneous gain for the MRT hybrid pre-coding. Here will not repeat it.

C. Effect of the Numbers of APs

Then it is able to obtain insight on the number of APs according to the semi-closed form approximation. It is noted that the increasing number of APs is reflected by the increasing values of M and $|\boldsymbol{\mathcal{A}}_k|$. Based on the semi-closed form, when the number of APs increases, it is able to find that the whole power consumption linearly increases, while the useful signal power increases in a second power function. When the transmit power is small, SINR is mainly affected by the useful signal power, and it is able to get that SINR increases with the number of APs in a second power function. Then the ergodic EE increases in this case. When the transmit power is large, SINR is both affected by the useful signal power and interference power. It means that SINR cannot increase with the increasing number of APs in a second power function, but the whole power consumption still linearly increases with the increasing number of APs. Then the ergodic EE decreases in this case. Therefore, based on the semi-closed form, it is able to get that the ergodic EE performance first increases and then decreases with the increasing number of APs, which are verified by results of Figs. 6 and 7.

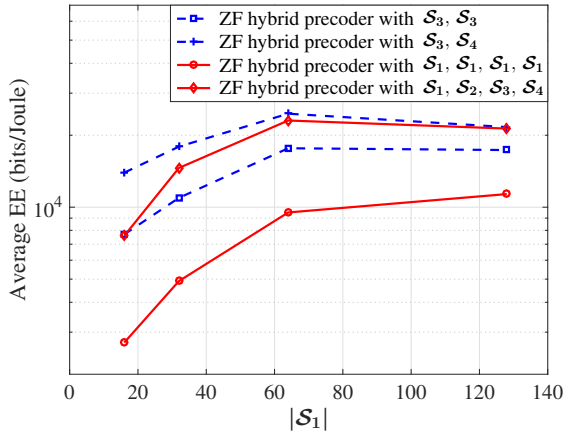


Fig. 6. Average EE performance with varying numbers of APs in the low SINR region. Here, $N_1 = 4$, and $P_1 = -10$ dBm.

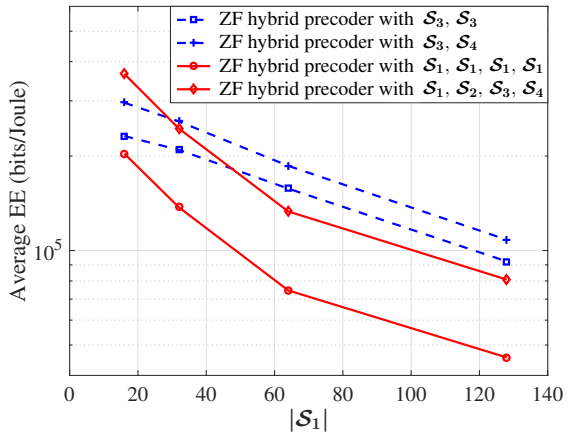


Fig. 7. Average EE performance with varying numbers of APs in the high SINR region. Here, $N_1 = 4$, and $P_1 = 20$ dBm.

Similarly, in the low SINR region, $\xi_{\mathrm{LW}}^{\mathrm{G,D},1}$ equals to $[0.2311, 0.2311, 0.2311, 0.2311]$ for the former group and $[0.5606, 0.5606, 0.5606, 0.5606]$ for the latter group, respectively. In the high SINR region, it was computed to be $[0.2253, 0.2253, 0.2253, 0.2253]$ for the former group and $[0.5573, 0.5573, 0.5573, 0.5573]$ for the latter

group. This phenomenon is found to be consistent with **Remark 1**, indicating the average EE gap of the former group is larger than that of the latter group. For the former and latter groups, $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ in the low SINR region was calculated to be $[1.9488, 1.8332, 2.1463, 1.9112]$ and $[1.1605, 1.3025, 1.2036, 1.2141]$, respectively, and $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ in the high SINR region was calculated to be $[1.4556, 1.4016, 1.2305, 1.2850]$ and $[1.0165, 1.1095, 1.0719, 1.1936]$, respectively. The values of $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ can be used to explain the simulation results in line with **Remark 2**. However, due to the relaxation of interference in the derivation of $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$, the difference between the former group and the latter group becomes small in the high SINR region, which does not agree well with the simulation results. This indicates that an accurate heterogeneous gain considering multiple interferences is required in future studies.

D. Effect of the Numbers of users

Based on the semi-closed form approximation, it is able to obtain insight on the number of users. When the number of users increases, limited by the number of RF chains in each AP and maximum transmit power of each AP, both the number of APs serving a user and the power allocated to a user decrease under the fair rule. Based on the semi-closed form, it is able to find that the whole power consumption linearly increases with the number of users. Besides, the useful signal power increases with the number of users in a linear function, but decreases in a second power function. Then the ergodic EE performance decreases in this case. Therefore, based on the semi-closed form, it is able to get that the ergodic EE decreases with the increasing number of users, as shown in Figs. 8 and 9.

In the low SINR region, we obtain the value of $\xi_{\mathrm{LW}}^{\mathrm{G,D},1}$, i.e., $[0.2311, 0.2267, 0.2245, 0.2233]$ for the former group and $[0.5606, 0.5581, 0.5568, 0.5562]$ for the latter group, respectively. In the high SINR region, it was computed to be $[0.2253, 0.2238, 0.2230, 0.2226]$ for the former group and $[0.5573, 0.5565, 0.5560, 0.5558]$ for the latter group. This phenomenon is found to be consistent with **Remark 1**, indicating the average EE gap of the former group is larger than that of the latter group. For the former and latter groups, $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ in the low SINR region was calculated to be $[3.2793, 3.5205, 3.8949, 3.6473]$ and $[2.1794, 2.2072, 2.4084, 2.3259]$, respectively, and $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ in the high SINR region was calculated to be $[2.4785, 2.0626, 2.0218, 1.3419]$ and $[1.2226, 1.1437, 1.4368, 1.1358]$, respectively. The values of $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ can be used to explain the simulation results in line with **Remark 2**.

E. Effect of the Numbers of Antennas

The effects of varying the numbers of transmit antennas on the ergodic EE in the low and high SINR regions are analyzed, respectively. As shown in Figs. 10 and 11, with the increasing number of transmit antennas, the average EE performances partly increase in the low SINR region while monotonously decreasing in the high SINR region. When the number of transmit antennas increases, the increasing interference in the high SINR region is larger than that in the low SINR region. This limits the sum rate performance and results in a monotonous decrease in the high SINR region.

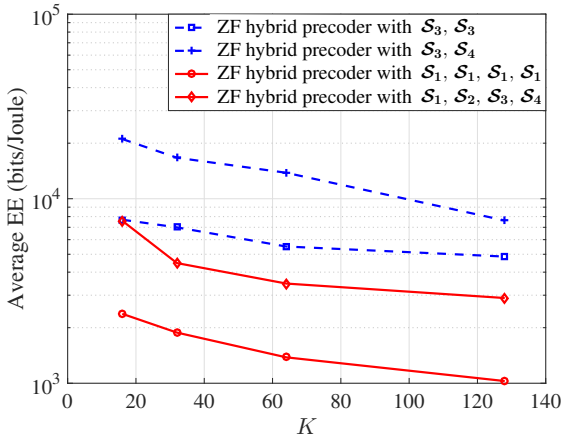


Fig. 8. Average EE performance with varying numbers of users in the low SINR region. Here, $N_1 = 4$, $|\boldsymbol{\mathcal{S}}_1| = 32$, and $P_1 = -10$ dBm.

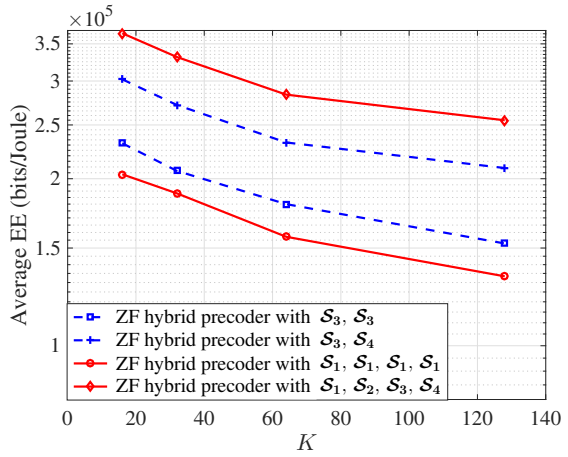


Fig. 9. Average EE performance with varying numbers of users in the high SINR region. Here, $N_1 = 4$, $|\boldsymbol{\mathcal{S}}_1| = 32$, and $P_1 = 20$ dBm.

In the low SINR regions, we computed $\xi_{\mathrm{LW}}^{\mathrm{G,D},1}$ as $[0.2311, 0.1815, 0.1542, 0.1398]$ for the former group and $[0.5606, 0.5323, 0.5167, 0.5085]$ for the latter group, whereas, in the high SINR region, we computed it to be $[0.2253, 0.1798, 0.1537, 0.1397]$ for the former group and $[0.5573, 0.5313, 0.5164, 0.5084]$ for the latter group. This analysis data explains why the average EE gap of the former group is larger than that of the latter group, in line with **Remark 1**. Further, we computed $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ for the former and latter groups to be $[6.8397, 3.3760, 4.2131, 2.6523]$ and $[1.1002, 1.6632, 1.7187, 1.4321]$ in the low SINR region, respectively, whereas, in the high SINR region, these were computed to be $[4.9308, 2.1846, 4.0879, 2.1530]$ and $[1.1536, 1.0218, 1.0089, 1.0214]$ for the former and latter groups, respectively. In line with **Remark 2**, the heterogeneous gain of the former group is found to be larger than that of the latter group, verified by the simulation results.

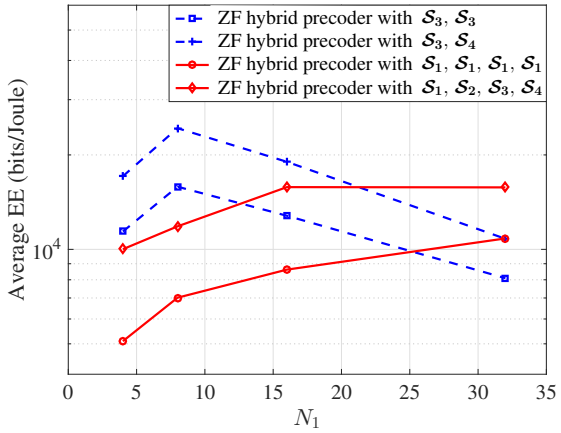


Fig. 10. Average EE performance with various numbers of transmit antennas in the low SINR region. Here, $|\boldsymbol{\mathcal{S}}_1| = 32$, and $P_1 = -10$ dBm.

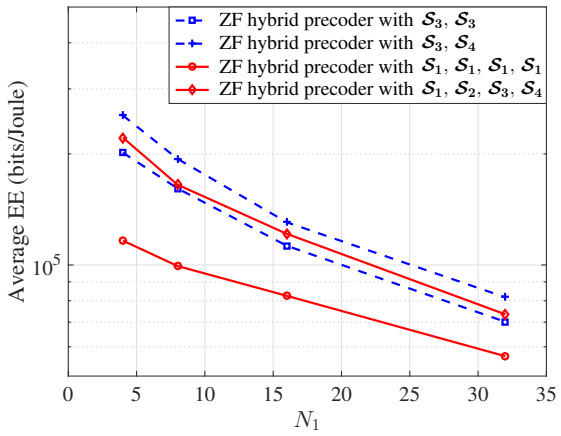


Fig. 11. Average EE performance with various numbers of transmit antennas in the high SINR power region. Here, $|\boldsymbol{\mathcal{S}}_1| = 32$, and $P_1 = 20$ dBm.

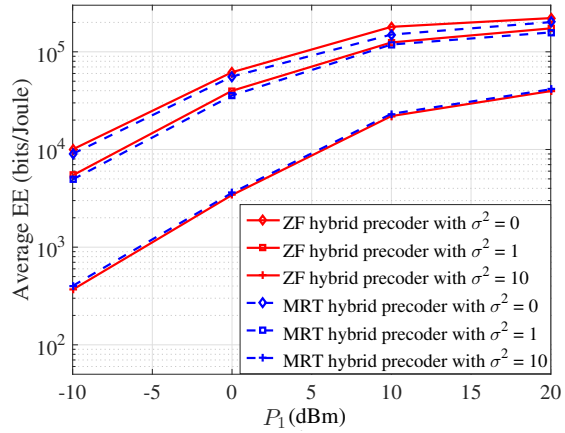


Fig. 12. Effect of imperfect CSI on the average EE with various transmit powers under $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_2, \boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$. Here, $|\boldsymbol{\mathcal{S}}_1| = 32$, and $N_1 = 4$.

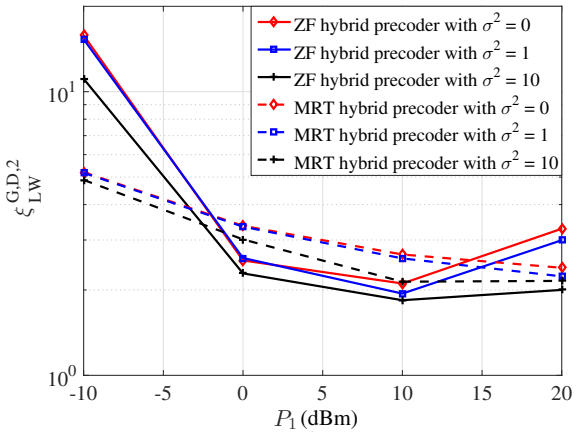


Fig. 13. Effect of imperfect CSI on $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ with various transmit powers under $\boldsymbol{\mathcal{S}}_1, \boldsymbol{\mathcal{S}}_2, \boldsymbol{\mathcal{S}}_3, \boldsymbol{\mathcal{S}}_4$. Here, $|\boldsymbol{\mathcal{S}}_1| = 32$, and $N_1 = 4$.

F. Effect of the Imperfect CSI

Finally, we research the effect of imperfect CSI on the ergodic EE of heterogeneous cell-free systems. In this case, the actual channel matrix between the mth AP and the kth user is expressed as $\mathbf{H}_{m,k} = \widehat{\mathbf{H}}_{m,k} + \widetilde{\mathbf{H}}_{m,k}$, where $\widehat{\mathbf{H}}_{m,k}$ is the estimated channel matrix. Each element of the estimation error matrix $\widetilde{\mathbf{H}}_{m,k}$ is modeled as a complex Gaussian random variable, with mean value being zero and normalized variance being σ^2. $\sigma^2 = 0$ corresponds to the perfect CSI. The number of channel realization is set as 1000, and discrete step equals to 0.01. Simulation results are shown in Fig. 12, with the increasing value of σ^2, which corresponds to serious imperfect CSI, the system performance becomes worse. So the semi-closed ergodic EE expression with the perfect CSI can be seen as an upper bound. Besides, when the CSI is perfect, ZF based hybrid pre-coding performs better than MRT based hybrid pre-coding, and the situation is opposite if the CSI is seriously imperfect, like $\sigma^2 = 10$. Compared to the ZF based hybrid pre-coding, the MRT based hybrid pre-coding is insensitive to the imperfect CSI.

However, a heterogeneous gain is a ratio of two ergodic EEs under various system parameters. Especially, the first case of heterogeneous gain is achieved without the complicated CSI and statistical information, which means $\xi_{\mathrm{LW}}^{\mathrm{G,D},1}$ is not affected by the imperfect CSI. The second case of heterogeneous gain is achieved by the semi-closed ergodic EE expression. With the same changed CSI, the changement of two ergodic EEs are similar. So the effect of imperfect CSI on $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ is weak. As shown in the simulation results of Fig. 13, the second cases of heterogeneous gain $\xi_{\mathrm{LW}}^{\mathrm{G,D},2}$ under both ZF and MRT based hybrid pre-codings change mildly, which indicates the weak effect of imperfect CSI on the heterogeneous gain.

VI. Conclusions

In this paper, we analyzed the ergodic EE of mmWave heterogeneous cell-free systems. Based on the system model, a general ergodic EE expression of mmWave heterogeneous cell-free systems was obtained. However, as this expression is not a closed form, it is quite difficult to extract useful information from it. We followed with a few relaxation steps to obtain a special semi-closed ergodic EE expression and

analyzed the effect of discretization, verifying the validity of this semi-closed ergodic EE expression. Based on this expression, we derived two cases of heterogeneous gain, describing the ergodic EE performance of mmWave heterogeneous cell-free systems: the first one describes the necessity of mmWave heterogeneous cell-free systems, and the second can accurately measure the heterogeneous gain and provide theoretical support for deploying heterogeneous networks. The simulation results demonstrated the effectiveness of the two cases of heterogeneous gain. Future work will focus on the ergodic EE of mmWave heterogeneous cell-free systems with multiple receive antennas at the UE side.

APPENDIX A
ERGODIC EE WITH A GENERAL HYBRID PRECODER

A generic ergodic EE with the statistical information is given in the manuscript. Based on it, a semi-closed ergodic EE is derived with several relaxing methods. Then two cases of heterogeneous gains are discussed. ZF based hybrid pre-coding works as an example to concretely illustrate corresponding processes. And these results can also apply to a more sophisticated hybrid pre-coding proposed in the literature. The reason lies in two aspects.

The first one is that a local ZF based hybrid pre-coding is taken for each AP, in order to trade off complexity and performance. This means that interferences between APs are not completely eliminated. So SINR still exists for each user in the downlink transmission. For a hybrid pre-coding proposed in the literature, such as MMSE and MRT based hybrid pre-codings, a similar SINR is achieved.

For a generic hybrid pre-coding scheme with $\mathbf{F}^{\mathrm{M}}$ and $\mathbf{W}^{\mathrm{M}}$ being the analog and digital parts, respectively, SINR of the kth user can be written as

$$\gamma_k^{\mathrm{M}} = \frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{P_{m,k}L_{m,k}}\mathbf{h}_{m,k}^{\mathrm{T}}\mathbf{F}_m^{\mathrm{M}}\mathbf{W}_{m,k}^{\mathrm{M}}|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2} \tag{36}$$

$P_k^{\mathrm{I}} = \sum_{k'\neq k}\sum_{m'\in\boldsymbol{\mathcal{A}}_{k'}}|\sqrt{P_{m',k'}L_{m',k'}}\mathbf{h}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}^{\mathrm{M}}\mathbf{W}_{m',k'}^{\mathrm{M}}|^2$. Recall the SINR expression under the ZF based hybrid pre-coding, the main difference lies in the molecule, that is $|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{P_{m,k}L_{m,k}}/\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\mathrm{F}}|^2$.

The second one is that a semi-closed expression of ergodic EE is achieved by three simplified steps, i.e., linearization of a PA efficiency, discretization of the integral, segmentation of a logarithmic function, which can apply to a situation taking a more sophisticated hybrid pre-coding. Here gives detailed explanations.

a) Linearization of a PA efficiency relaxes the nonlinear PA efficiency to a linear one. Since the power of a hybrid pre-coding scheme is normalized, the variable part of the sum power consumption, caused by PAs, is decided by the power allocation matrix. So for various hybrid pre-coding schemes, the expression of the variable part of the sum power consumption is the same, i.e.,

$$P_{\mathrm{V}}^{\mathrm{M}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}) \approx \sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\epsilon\left(|\boldsymbol{\mathcal{A}}_m^{-1}|-1\right)\left(\sum_{k\in\boldsymbol{\mathcal{A}}_m^{-1}}P_{m,k}\rho_t\right), \tag{37}$$

where $P_{m,k}$ is the power allocated to the kth user at the mth AP.

b) Discretization of the integral relaxes the calculation of SINR's PDF. Since the SINR expressions are similar under various hybrid pre-coding schemes, discretization steps and results are also similar. Then discrete ergodic EE can be expressed as,

$$\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,M}}(\boldsymbol{\mathcal{S}}) = B\sum_{i=1}^{I}\sum_{i'=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{L}\sum_{o=1}^{O}\sum_{o'=1}^{O}p(i,i',j,l,o,o') \frac{\sum_{k=1}^{K}\log_2\left(1+\frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{L_{m,k}^{o}P_{m,k}^{i}}\eta_{m,k}^{j}|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2}\right)}{P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})} \tag{38}$$

with

$$P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}) \approx \sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\epsilon\left(|\boldsymbol{\mathcal{A}}_m^{-1}|-1\right)\Big(P_{\mathrm{BB}}+ |\boldsymbol{\mathcal{A}}_m^{-1}|P_{\mathrm{RF}}+|\boldsymbol{\mathcal{A}}_m^{-1}|N_tP_{\mathrm{PS}}+\sum_{k\in\boldsymbol{\mathcal{A}}_m^{-1}}\sum_{i=1}^{I}\delta_{m,k}^{i}P_{m,k}^{i}\rho_t\Big). \tag{39}$$

The interference item P_k^{I} equals to $\sum_{m'\notin\boldsymbol{\mathcal{A}}_k}\sum_{k'\in\boldsymbol{\mathcal{A}}_{m'}^{-1}}(\sum_{i'=1}^{I}\delta_{m,k}^{i'}\sqrt{P_{m',k'}^{i'}})^2 \quad (\sum_{o'=1}^{O}\delta_{m',k'}^{o'}\sqrt{L_{m',k'}^{o'}})^2 (\sum_{l=1}^{L}\delta_{m',k',k}^{l}\mu_{m',k',k}^{l})$. $L_{m,k}^{o}$ ($L_{m',k'}^{o'}$), $P_{m,k}^{i}$ ($P_{m',k'}^{i'}$), $\eta_{m,k}^{j}$, and $\mu_{m',k',k}^{l}$ are the discrete $L_{m,k}$ ($L_{m',k'}$), $P_{m,k}$ ($P_{m',k'}$), $\mathbf{h}_{m,k}^{\mathrm{T}}\mathbf{F}_m^{\mathrm{M}}\mathbf{W}_{m,k}^{\mathrm{M}}$, and $\mathbf{h}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}^{\mathrm{M}}\mathbf{W}_{m',k'}^{\mathrm{M}}$, respectively.

c) Segmentation of a logarithmic function relaxes the sum of achievable rates. Since the segmentation of a logarithmic function is given, i.e., $\log_2(1+\gamma_k)\approx\beta_s\gamma_k+C_s$, it is the same for various hybrid pre-coding schemes.

Therefore, the semi-closed expression of ergodic EE can still come into existence with a minor adjustment under various hybrid pre-codings, which is

$$\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,M}}(\boldsymbol{\mathcal{S}}) \approx B\sum_{i=1}^{I}\sum_{i'=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{L}\sum_{o=1}^{O}\sum_{o'=1}^{O}p(i,i',j,l,o,o') \frac{\sum_{k=1}^{K}\sum_{s=1}^{S}\left(\beta_s\frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{L_{m,k}^{o}P_{m,k}^{i}}\eta_{m,k}^{j}|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2}+C_s\right)\delta_{k,s}}{P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})} \tag{40}$$

where $\delta_{k,s}=1$ when γ_k lies in the sth segment. Otherwise, $\delta_{k,s}=0$.

Considering that two versions of heterogeneous gains are achieved based on the semi-closed expression of ergodic EE, they also exist for various hybrid pre-coding schemes. Especially, the first case of heterogeneous gain ($\xi_{\mathrm{LW}}^{\mathrm{G,D,1}}$) is unrelated to hybrid pre-coding and stays the same under various hybrid pre-coding schemes. The second case of heterogeneous gain is the ratio of two lower bounds of semi-closed expressions, that is $\xi_{\mathrm{LW}}^{\mathrm{G,D,2}}=\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}}_{\max})/\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,LW}}(\boldsymbol{\mathcal{S}}_{\min})$. Once the semi-closed expression of ergodic EE under a general hybrid pre-coding is given, it is easy to have $\xi_{\mathrm{LW}}^{\mathrm{G,D,2}}$.

APPENDIX B
ERGODIC EE UNDER CHANNELS WITH MULTIPLE PATHS

The front analyses on ergodic EE are conducted on the LoS approximation, where the sum of LoS fading and NLoS

$$\gamma_k^{\mathrm{MP}} = \frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{P_{m,k}L_{m,k}}/\|\mathbf{F}_m\widetilde{\mathbf{W}}_{m,k}\|_{\mathrm{F}} + \sqrt{P_{m,k}}(\mathbf{H}_{m,k}^{\mathrm{NS}})^{\mathrm{T}}\mathbf{F}_m\mathbf{W}_{m,k}|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2} \tag{43}$$

$$\begin{aligned}\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,MP}}(\boldsymbol{\mathcal{S}}) \approx & B\sum_{i=1}^{I}\sum_{i'=1}^{I}\sum_{j=1}^{J}\sum_{j'=1}^{J}\sum_{l=1}^{L}\sum_{l'=1}^{L}\sum_{o=1}^{O}\sum_{o'=1}^{O} p\left(i,i',j,j',l,l',o,o'\right) \\ & \frac{\sum_{k=1}^{K}\sum_{s=1}^{S}\left(\beta_s\frac{|\sum_{m\in\boldsymbol{\mathcal{A}}_k}\sqrt{L_{m,k}^{o}P_{m,k}^{i}}\eta_{m,k}^{j}+\sqrt{P_{m,k}^{i}}\eta_{m,k}^{j'}|^2}{P_k^{\mathrm{I}}+\sigma_{\mathrm{noise}}^2}+C_s\right)\delta_{k,s}}{P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}})}\end{aligned} \tag{44}$$

fading is replaced by LoS fading. When the channel gain of LoS path is obviously larger than that of NLoS path, the LoS approximation is tenable. It is noted that when the number of paths is larger than one, or the power difference between various paths is small, our work still comes into existence and similar analyses can be achieved. To elaborate these, two processing methods are considered as follows.

a) In the first case, the channel matrix between the mth AP and kth user is modeled as

$$\mathbf{H}_{m,k} = \gamma\sum_{i=1}^{N_{\mathrm{cl}}}\sum_{l=1}^{N_{\mathrm{ray},i}}\alpha_{i,l}\sqrt{L_{i,l}}\mathbf{a}\left(\theta_{i,l}\right), \tag{41}$$

where γ is the normalized factor. N_{cl} is the number of scattering clusters, each of which is consisted of $N_{\mathrm{ray},i}$ propagation paths. $\alpha_{i,l}\sim\mathcal{CN}(0,\sigma_{i,l}^2)$ and $L_{i,l}$ are the small-scale fading and the large-scale fading, respectively. $\mathbf{a}\left(\theta_{i,l}\right)$ represents the normalized transmit array response vector evaluated at the corresponding angles of departure, that is $\theta_{i,l}$.

Since the small-scale fading follows the complex Gaussian distribution with zero mean value, it is reasonable to treat the mean value of the sum fading of multiple paths as one large-scale fading. Then the multiple paths can be seen as one equivalent path, with equivalent large-scale fading being the mean value of the sum fading and equivalent small-scale fading being a complex Gaussian random variable. Hybrid pre-coding scheme is conducted under this equivalent path, then the proposed analyses in the manuscript can be directly employed.

b) In the second case, the channel matrix is modeled as

$$\mathbf{H}_{m,k} = \mathbf{H}_{m,k}^{\mathrm{S}} + \mathbf{H}_{m,k}^{\mathrm{NS}}, \tag{42}$$

where $\mathbf{H}_{m,k}^{\mathrm{S}}$ is the path with the maximum channel gain, and $\mathbf{H}_{m,k}^{\mathrm{NS}}$ is the sum of other paths.

Hybrid pre-coding scheme is still only conducted under the path with the maximum channel gain. Under this situation, SINR of the kth user is (43), where $P_k^{\mathrm{I}} = \sum_{k'\neq k}\sum_{m'\in\boldsymbol{\mathcal{A}}_{k'}}|\sqrt{P_{m',k'}L_{m',k'}}\mathbf{h}_{m',k}^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'} + \sqrt{P_{m',k'}}(\mathbf{H}_{m',k}^{\mathrm{NS}})^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}|^2$. Compared to the SINR γ_k in (13) with only LoS, the main differences lie in the additional items in the molecule and denominator of γ_k^{MP}.

Based on SINR expression of γ_k^{MP}, it is able to get the semi-closed expression of ergodic EE by three main simplified steps, i.e., linearization of a PA efficiency, discretization of the integral, segmentation of a logarithmic function. Since linearization of a PA efficiency is decided by the power allocation matrix, and segmentation of a logarithmic function is decided by value of SINR, these two simplified steps are same to the case with only the LoS. For the discretization of the integral, it is necessary to discrete the added items, that is $\sqrt{P_{m,k}}(\mathbf{H}_{m,k}^{\mathrm{NS}})^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}$ and $\sqrt{P_{m',k'}}(\mathbf{H}_{m',k}^{\mathrm{NS}})^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}$.

Then the semi-closed expression of ergodic EE under channels with multiple paths becomes (44), where $P_{\mathrm{all}}^{\mathrm{C}}(\boldsymbol{\mathcal{H}}|\boldsymbol{\mathcal{S}}) = \sum_{m=1,m\in\boldsymbol{\mathcal{S}}_t}^{M}\epsilon(|\boldsymbol{\mathcal{A}}_m^{-1}|-1)\times(P_{\mathrm{BB}}+|\boldsymbol{\mathcal{A}}_m^{-1}|P_{\mathrm{RF}}+|\boldsymbol{\mathcal{A}}_m^{-1}|N_tP_{\mathrm{PS}}+\sum_{k\in\boldsymbol{\mathcal{A}}_m^{-1}}P_{m,k}^{i}\rho_t)$ and P_k^{I} is equal to $\sum_{m'\notin\boldsymbol{\mathcal{A}}_k}\sum_{k'\in\boldsymbol{\mathcal{A}}_{m'}^{-1}}\left(P_{m',k'}^{i'}L_{m',k'}^{o'}\mu_{m',k',k}^{l}+P_{m',k'}^{i'}\mu_{m',k',k}^{l'}\right)$. $\eta_{m,k}^{j'}$ and $\mu_{m',k',k}^{l'}$ are the discretized values of $(\mathbf{H}_{m,k}^{\mathrm{NS}})^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}$ and $(\mathbf{H}_{m',k}^{\mathrm{NS}})^{\mathrm{T}}\mathbf{F}_{m'}\mathbf{W}_{m',k'}$.

Based on the semi-closed expression of ergodic EE, two cases of heterogeneous gains are achieved. With $\bar{\xi}_{\mathrm{EE}}^{\mathrm{D,MP}}(\boldsymbol{\mathcal{S}})$, it is able to get the corresponding two versions of heterogeneous gains by the same way to utilize $\bar{\xi}_{\mathrm{EE}}^{\mathrm{D}}(\boldsymbol{\mathcal{S}})$ with only LoS path.

REFERENCES

[1] S. Sun, T. S. Rappaport, M. Shafi, P. Tang, J. Zhang, and P. J. Smith, "Propagation models and performance evaluation for 5G millimeter-wave bands," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8422-8439, 2018.

[2] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE Access* , vol. 8, pp. 133995-134030, 2020.

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834-1850, 2017.

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

[5] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 706-709, Dec. 2017.

[6] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1250-1264, Feb. 2020.

[7] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, 2019, pp. 1-6.

[8] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247-4261, July 2020.

[9] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity," *IEEE Access*, vol. 7, pp. 44596-44612, 2019.

[10] T. Zhao, S. Chen, R. Zhang, H. -H. Chen, and Q. Guo, "Uplink channel estimation with reduced fronthaul overhead in cell-free massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 11, no. 8, pp. 1718-1722, Aug. 2022.

[11] Y. Li and G. A. Aruma Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 7, no. 6, pp. 950-953, Dec. 2018.

[12] R. Sayyari, J. Pourrostam, and M. J. M. Niya, "Cell-free massive MIMO system with an adaptive switching algorithm between cooperative NOMA, non-cooperative NOMA, and OMA modes," *IEEE Access*, vol. 9, pp. 149227-149239, 2021.

[13] Z. Yang and Y. Zhang, "Beamforming optimization for RIS-aided SWIPT in cell-free MIMO networks," *China Communications*, vol. 18, no. 9, pp. 175-191, Sept. 2021.

[14] Y. Zhang *et al.*, "Beyond cell-free MIMO: Energy efficient reconfigurable intelligent surface aided cell-free MIMO communications," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 2, pp. 412-426, June 2021.

[15] C. D'Andrea, A. Garcia-Rodriguez, G. Geraci, L. G. Giordano, and S. Buzzi, "Analysis of UAV communications in cell-free massive MIMO systems," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 133-147, 2020.

[16] J. Zheng, J. Zhang, and B. Ai, "UAV communications with WPT-aided cell-free massive MIMO systems," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3114-3128, Oct. 2021.

[17] A. Ghosh *et al.*, "Heterogeneous cellular networks: From theory to practice," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 54-64, June 2012.

[18] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20-27, June 2013.

[19] S. Kim and B. Shim, "Energy-efficient millimeter-wave cell-free systems under limited feedback," *IEEE Transactions on Communications*, vol. 69, no. 6, pp. 4067-4082, June 2021.

[20] A. Liu and V. K. N. Lau, "Joint BS-user association, power allocation, and user-side interference cancellation in cell-free heterogeneous networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 335-345, 15 Jan.15, 2017.

[21] J. An and F. Zhao, "Trajectory optimization and power allocation algorithm in MBS-assisted cell-free massive MIMO systems," *IEEE Access*, vol. 9, pp. 30417-30425, 2021.

[22] Z. Niu, W. Ma, W. Wang, and T. Jiang, "Spatial modulation-based ambient backscatter: Bringing energy self-sustainability to massive internet of everything in 6G," *China Communications*, vol. 17, no. 12, pp. 52-65, Dec. 2020.

[23] S. Shen, C. Yu, K. Zhang, J. Ni, and S. Ci, "Adaptive and dynamic security in AI-empowered 6G: From an energy efficiency perspective," *IEEE Communications Standards Magazine*, vol. 5, no. 3, pp. 80-88, September 2021.

[24] D. Feng, C. Jiang, G. Lim, L. J. Cimini, G. Feng, and G. Y. Li, "A survey of energy-efficient wireless communications," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 167-178, First Quarter 2013.

[25] X. Gao, L. Dai, S. Han, C. -L. I, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 998-1009, April 2016.

[26] J. -C. Guo, Q. -Y. Yu, W. -X. Meng, and W. Xiang, "Energy-efficient hybrid precoder with adaptive overlapped subarrays for large-array mmWave systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1484-1502, March 2020.

[27] R. Hamdi and M. Qaraqe, "Energy cooperation in renewable-powered cell-free massive MIMO systems," in *Proc. 2019 25th Asia-Pacific Conference on Communications (APCC)*, Ho Chi Minh City, Vietnam, 2019, pp. 305-309.

[28] T. C. Mai, H. Q. Ngo, and L. -N. Tran, "Energy efficiency maximization in large-scale cell-free massive MIMO: A projected gradient approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6357-6371, Aug. 2022.

[29] H. Q. Ngo, L. -N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 1, pp. 25-39, March 2018.

[30] J. García-Morales, G. Femenias, and F. Riera-Palou, "Energy-efficient access-point sleep-mode techniques for cell-free mmWave massive MIMO networks with non-uniform spatial traffic density," *IEEE Access*, vol. 8, pp. 137587-137605, 2020.

[31] S. Chen, R. Ma, H. -H. Chen, H. Zhang, W. Meng, and J. Liu, "Machine-to-machine communications in ultra-dense networks—A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1478-1503, third quarter 2017.

[32] S. Chen, X. Liu, T. Zhao, H. -H. Chen, and W. Meng, "Performance analysis of joint transmission schemes in ultra-dense networks – A unified approach," *IEEE/ACM Transactions on Networking*, vol. 28, no. 1, pp. 154-167, Feb. 2020.

[33] S. Chen, T. Zhao, H. H. Chen, and W. Meng, "Network densification and path-loss models versus UDN performance—A unified approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4058-4071, July 2021.

[34] S. Elhoushy, M. Ibrahim, and W. Hamouda, "Cell-free massive MIMO: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 492-523, First quarter 2022.

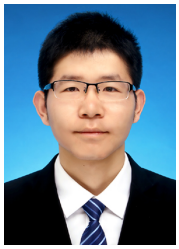
[35] J. C. Guo, Q. Y. Yu, W. X. Meng, and W. Xiang, "Ergodic energy efficiency of mmWave system considering insertion loss under dynamic subarray architecture," in *Proc. 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, Antwerp, Belgium, 2020, pp. 1-5.

[36] T. S. Rappaport, G. R. Maccartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.

[37] M. R. Akdeniz, *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.

[38] S. Buzzi, C. D'Andrea, M. Fresia, and X. Wu, "Multi-UE multi-AP beam alignment in user-centric cell-free massive MIMO systems operating at mmWave," *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 8919-8934, Nov. 2022.

[39] Y. Zhang, Z. Cai, and G. Xiong, "A new image compression algorithm based on non-uniform partition and u-system," *IEEE Transactions on Multimedia*, vol. 23, pp. 1069-1082, 2021.



Jichong Guo (Member, IEEE) received the M.S. and Ph.D. degrees from Harbin Institute of Technology (HIT), P. R. China, in 2016 and 2021, respectively. From 2018 to 2019, he was a visiting student in the Department of Electrical and Computer Engineering, The University of British Columbia (UBC), Vancouver, B.C., Canada. Currently he is a lecturer in School of Electronic and Information Engineering, Suzhou University of Science and Technology. His research interests include wireless channel modeling and pre-coding technology.

Dekun Zhang received the B.S., M.S. degrees in communications engineering from Harbin Institute of Technology (HIT), P.R.China, in 2013, 2015 respectively. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Integrated Services Networks, Xidian University, P. R. China. And he is the chief algorithm expert of ZTE Communications Co., Ltd. His research interests include signal processing for wireless communications and large-scale distributed MIMO systems (cellfree massive MIMO).

Chen Cui (Member, IEEE) received the B.S. degree in communication engineering, and the M.S. and Ph.D. degrees in electronics and communication engineering from the Harbin Institute of Technology, Harbin, China, in 2013, 2015, and 2022, respectively. She is a Post Doctoral Researcher with the School of Computer Science, Peking University. Her research interests include source and channel coding and image/video transmission.

Xiqing Liu (Member, IEEE) received his M.Sc. and Ph.D. degrees from Harbin University of Science and Technology and Harbin Institute of Technology, Harbin, China, in 2012 and 2017, respectively. From 2018 to 2019, he was a postdoctoral fellow with the Department of Engineering Science, National Cheng Kung University, Taiwan. Currently, he is an associate professor in State Key Laboratory of Networking and Switching Technology (SKLNST), School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT). His current research interests include interference suppression in multicarrier systems, non-orthogonal multiple access and MIMO technologies.