This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3611133

Multiple CPUs Cooperation for CF Massive MIMO With MmWave Fronthaul and Backhaul

Feiyang Li, Qiang Sun, *Member, IEEE*, Jiayi Zhang, *Senior Member, IEEE*, Cunhua Pan, *Senior Member, IEEE*, and Kai-Kit Wong, *Fellow, IEEE*

***Abstract*—Cell-free massive multiple-input multiple-output (CF massive MIMO) is regarded as a promising technology for next-generation wireless communication systems. However, relying on a single central processing unit (CPU) in CF massive MIMO systems is not scalable in practical networks, requiring the introduction of multiple CPUs for more efficient and feasible transmission. In this paper, we investigate a CF massive MIMO system with multiple CPUs. To obtain flexible and cost-efficient deployment, we propose to use wireless x-haul links instead of wired ones. More specifically, we assume that both the fronthaul links from the APs to the corresponding CPU and the backhaul links between CPUs operate under millimeter wave (mmWave) networks. Taking into account a tradeoff between the degree of centralized coordination and the signal overhead on the backhaul links, we consider four levels of multiple CPUs cooperation schemes from fully centralized to fully distributed. In addition, we propose a binary search method to allocate the backhaul capacities for maximizing the sum spectral efficiency (SE). Simulation results show that mmWave backhaul amplifies the compression noise introduced by mmWave fronthaul, leading to a more pronounced impact on the SE of systems. In this case, the centralized processing scheme can generate more compression noise due to the larger data overhead on the backhaul link, making the distributed processing scheme a superior processing scheme, especially when dealing with a large number of APs or significant distances between CPUs.**

***Index Terms*—Cell-free massive MIMO, multiple CPUs cooperation, mmWave fronthaul and backhaul, spectral efficiency.**

I. Introduction

Cell-free massive multiple-input multiple-output (CF massive MIMO) is viewed as a prospective technology for future wireless networks [1], [2]. In CF massive MIMO networks, numerous access points (APs) are geographically distributed across the coverage area, connecting to a central processing unit (CPU) via fronthaul links, and collectively provide coherent service to multiple user equipments (UEs) [3].

Similar to cellular massive MIMO, CF massive MIMO can employ the favorable propagation and channel hardening properties when the number of APs is large to multiplex many UEs sharing the same time-frequency resource [4], [5]. Therefore, it can provide high spectral efficiency (SE) with the usage of simple signal processing [6], [7]. Nowadays, several works on CF massive MIMO focused on the maximum ratio (MR) processing [8], [9], while the work reported in [10] found that higher SE is achieved when the minimum mean-square error (MMSE) processing is applied. In particular, by the CF massive MIMO configuration, the APs are placed close to the UEs, which yields a high macro-diversity and low path-loss fading [11]. As a result, numerous UEs can be served simultaneously with uniformly good quality-of-service [12].

Most existing papers focused on CF massive MIMO systems with a single CPU. However, relying on a single CPU is not scalable due to factors such as signal processing complexity and transmission overhead in practical networks. To this end, some researchers explored the introduction of multiple CPUs to facilitate more efficient and feasible transmission. More specifically, the authors of [13] first proposed multiple CPUs to achieve the scalability in CF massive MIMO systems. A mixed coherent and non-coherent transmission scheme was considered in [14] and [15], while [16] introduced a hybrid configuration for virtualized CPUs to improve wireless communication quality and throughput. Notably, multiple CPU cooperation schemes gained significant attention for enabling joint signal processing and enhancing system performance. In [17], the authors demonstrated that systems without CPUs cooperation experience a 28.66% performance degradation compared to those with cooperation. [18] proposed a scalable CPU cooperation scheme that focuses on power control to address the performance degradation issue in the CPU edge region. Despite these efforts, the current understanding of CPU cooperation in CF massive MIMO systems remains incomplete, lacking comprehensive comparison and analysis across different levels of CPU cooperation. In addition, all of these works assumed error-free information transmission between CPUs, which is unfeasible for practical systems, further highlighting the need for more accurate and practical evaluations.

Moreover, x-haul links represent the networks that transmit the signals between the APs and the CPUs in CF massive MIMO, which can significantly affects the system performance [19]. In fact, the limitations of the x-haul link between APs

This work was supported in part by the National Natural Science Foundation of China under Grant 62371262, Grant 62401297, and Grant 62341131, in part by the Qinlan Project of Jiangsu Province, in part by the Nantong Key Research and Development Program under Grant GZ2024002 and in part by the Natural Science Foundation of Nantong under Grant JC2023018 (*Corresponding author: Qiang Sun.*)

F. Li and Q. Sun are with the School of Information Science and Technology, Nantong University, Nantong 226019, China. (e-mail: lfy@stmail.ntu.edu.cn; sunqiang@ntu.edu.cn).

J. Zhang is with the School of Electronic and Information Engineering and the Frontiers Science Center for Smart High-speed Railway System, Beijing Jiaotong University, Beijing 100044, China. (e-mail: jiayizhang@bjtu.edu.cn).

Cunhua Pan is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. (e-mail: cpan@seu.edu.cn)

K. K. Wong is affiliated with the Department of Electronic and Electrical Engineering, University College London, Torrington Place, WC1E 7JE, United Kingdom and he is also affiliated with Yonsei Frontier Lab, Yonsei University, Seoul, Korea. (email: kitwong@ieee.org)

and the CPU, i.e., limited fronthaul link, have been widely investigated [20]–[24]. For instance, [20] and [21] analyzed the SE of CF massive MIMO systems with constrained fronthaul, concluding that compression noise resulting from limited fronthaul degrades the SE of system. the authors of [22] further demonstrated that centralized operation outperforms distributed operation even though more compression noise is generated. In addition, employing optimization algorithms for fronthaul capacity allocation was regarded as an effective method for mitigating the negative effects of limited fronthaul links [23], [24]. However, all of these works are built on the wire-based fronthaul networks, which constrain the system scalability as the network expands, i.e., the coverage area increases or the number of APs grows. This limitation arises because cables or fibers are not always readily available in many urban locations. As a result, implementing a wireless fronthaul network presents a more practical and scalable solution compared to its wired counterpart, offering greater flexibility and cost-efficient deployment [25].

Recently, wireless fronthaul networks attracted much attention from both academia and industry [26]. The authors of [27] first proposed a CF massive MIMO architecture utilizing higher-band fronthaul, e.g., a millimeter wave (mmWave) or terahertz (THz) fronthaul, for sub-6GHz systems. [35] revealed that the centralized operation demonstrates superior performance in comparison to the distributed operation in mmWave networks, particularly when the APs were equipped with decoding capabilities. Furthermore, the SINR with MMSE combining scheme for different fronthaul schemes was provided in [29]. However, the aforementioned studies only considered the mmWave fronthaul from APs to the CPU [26], [27], [29], [35]. In multi-CPU systems, since the distances between CPUs are typically very long, the cost of deploying fibers or cables may be significantly increased due to signal attenuation necessitating additional equipment and the complexity of installation across difficult terrains. Thus, the usage of mmWave for the x-haul links (which can be called "mmWave backhaul link") is a more cost-effective solution [27]. Nowadays, few works have explored the utilization of mmWave for backhaul. Though the authors considered the mmWave backhaul in [30], they disregarded the cooperation between multiple CPUs, instead adopting a non-cooperative signal processing approach. This implies that [30] may underestimate the performance of mmWave backhaul networks. Thus, it is imperative to build a comprehensive analysis framework of CF massive MIMO with mmWave backhaul networks.

Motivated by the above observations, we consider a CF massive MIMO system with multiple CPUs. We take into account the usage of mmWave for the x-haul links. In particular, the backhaul link between CPUs is also applied by mmWave communication. Moreover, we consider four levels of multiple CPUs cooperation based on the signal overhead of the backhaul links. Finally, we propose a binary search method to allocate the backhaul capacities for maximizing the sum SE. The major contributions of this paper are listed as follows:

1) We investigate a CF massive MIMO system with multiple CPUs, where mmWave is utilized for the x-haul links. More specifically, both the fronthaul links from the APs to the corresponding CPU and the backhaul links between CPUs operate under mmWave networks. Furthermore, we also introduce rate-distortion theory for mmWave fronthaul and backhaul links.
2) Taking into account a tradeoff between the degree of centralized coordination and the signal overhead on the backhaul links, we consider four multi-CPU cooperation schemes from fully centralized to fully distributed inspired by [10]. Moreover, we derive novel closed-form SE expressions for Level 2 and Level 3 using the MR combining. In addition, we propose a binary search method to allocate the backhaul capacities for maximizing the sum SE.
3) Through the simulation results, we observe mmWave backhaul amplifies the compression noise introduced by the fronthaul, leading to a more pronounced impact on the SE of systems. Nonetheless, the proposed CPU cooperation schemes still outperform the non-cooperative scheme, with the performance gains increasing as the number of CPUs grows. Besides, as the number of APs and the distance between CPUs increase, the fully centralized processing scheme loses its advantage due to the intensified compression noise, making the distributed processing scheme with large-scale fading decoding (LSFD) a more competitive option under these conditions.

The rest of this paper is organized as follows. In Section II, we propose and describe the system model of CF massive MIMO with multiple CPUs. Next, we present the signal processing schemes within CPUs in Section III. Then, four levels of multi-CPU cooperation schemes are proposed in Section IV. In Section V, simulation results show the impact of mmWave fronthaul and backhaul, comparing the SE performance across various CPU cooperation levels. Finally, the conclusion of this paper is presented in Section VI.

Notation: Bold lowercase letters denote column vectors, while bold uppercase letters signify matrices. Superscripts $(\cdot)^*, (\cdot)^T$, and $(\cdot)^H$ indicate conjugate, transpose, and conjugate transpose, respectively. Symbol $\triangleq$ is employed for definitions, and the identity matrix of size $N \times N$ is showed as $\mathbf{I}_N$. We use notation $\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$ to denote the multi-variate circularly symmetric complex Gaussian distribution with a correlation matrix $\mathbf{R}$. The expected value of random variable $\mathbf{g}$ is represented as $\mathbb{E}\{\mathbf{g}\}$.

II. System Model

A. Network Architecture

The proposed CF massive MIMO system is illustrated in Fig. 1, in which multiple CPUs are placed to control all APs for transmission. More specifically, J CPUs, $J \times L$ APs, and K single-antenna UEs are considered in the network. Each AP is equipped with N antennas. We assume that each CPU is connected to L APs via mmWave fronthaul. In particular, all these CPUs are also connected to each other through mmWave backhaul. In CPU j, the channel between AP l and UE k is expressed as

$$\mathbf{h}_{jlk} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}_{jlk}\right), \tag{1}$$

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3611133

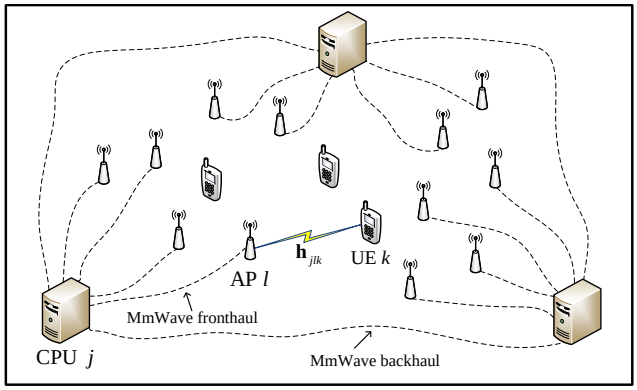


Fig. 1: CF massive MIMO with multiple CPUs system model.

where $\mathbf{R}_{jlk} \in \mathbb{C}^{N\times N}$ is the spatial correlation matrix, which characterizes the spatial properties of the channel and $\beta_{jlk} \triangleq \mathrm{tr}\left(\mathbf{R}_{jlk}\right)/N$ symbolizes the large-scale fading coefficient related to geometric path-loss and shadowing.

The time-division duplex (TDD) protocol is applied, where all APs serve all UEs using the same time-frequency resource. In addition, this paper considers the uplink, which consists of τ_p channel uses dedicated for pilots and $\tau_c - \tau_p$ channel uses for payload data.

B. Achievable Rate for Fronthaul and Backhaul

MmWave frequency band offers the advantages of the large available bandwidth and the high beamforming gains, making it suitable for supporting high-speed wireless fronthaul and backhaul for different levels of cooperation. Since the distance between the CPUs and their associated APs, as well as the distance between the CPUs themselves, is typically very long, the existence of line-of-sight (LoS) links cannot be guaranteed. Therefore, we aspire to exam the systems under non-line-of-sight (NLoS) conditions to better reflect the scenarios in practice. The reasons for using the sub-6GHz band for the access links and the mmWave band for the x-haul links are as follows: (i) The natural frequency separation between the sub-6GHz links and the higher mmWave links effectively mitigates the interference, simplifying systems resource scheduling and RF design. (ii) The shorter wavelength of the higher-frequency fronthaul signals enables high-precision synchronization across the APs with sub-6GHz, which is crucial for coherent joint processing [27]. Next, we present the path-loss fading model using the link between the CPUs and their associated APs as an example. From [27], [31], the reference distance is set to 1 m, and the transmission power P_f with propagation distance r_l of AP l is attenuated by factor r_l^α where α is the path-loss exponent. Thus, the corresponding signal-to-noise ratio (SNR) of the AP l can be written as

$$\gamma_l = \frac{P_f G_t}{A r_l^\alpha N_0 B}, \tag{2}$$

where N_0 represents the thermal noise power per Hz, A characterizes the path-loss intercept, G_t is the total gain for the link between the CPU and the associated APs, and B denotes the link bandwidth.

After determining the SNR for each AP, we need to calculate the total fronthaul capacity allocated to all APs. For this purpose, it is assumed that the available fronthaul bandwidth is divided equally between these APs and is reused between the CPUs with a factor of one. Therefore, the assigned bandwidth to each AP is

$$B_{f,l} = \chi_l B_f, \tag{3}$$

where χ_l is the bandwidth distribution (BD) factor for AP l. Considering the noise-limited system for mmWave communication, the fronthaul capacity for the link between AP l and its associated CPU j can be defined as

$$\begin{aligned} C_{jl}^* &= B_{f,l} \times \log_2(1+\gamma_l) \\ &= \chi_l B_f \times \log_2\left(1 + \frac{P_f G_t}{A r_l^\alpha N_0 \chi_l B_f}\right). \end{aligned} \tag{4}$$

To assess how mmWave fronthaul link capacities affect the system performance, it is essential to normalize the capacity of these links relative to the access link bandwidth. Since all UEs utilize the full available access bandwidth, the normalized mmWave fronthaul capacity for access point l linked to CPU j can be expressed as

$$C_{f,jl} = \frac{C_{jl}^*}{B_A} = \frac{\chi_l B_f}{B_A} \times \log_2\left(1 + \frac{P_f G_t}{A r_l^\alpha N_0 \chi_l B_f}\right). \tag{5}$$

Without loss of generality, we assume that the mmWave backhaul link capacities between the CPUs also follow the above setups [27]. Thus, backhaul capacity of CPU j is given by

$$C_{b,j} = \frac{C_j^*}{B_A} = \frac{\chi_j B_b}{B_A} \times \log_2\left(1 + \frac{P_b G_t}{A r_j^\alpha N_0 \chi_j B_b}\right), \tag{6}$$

where χ_j is the BD factor for AP j.

Remark 1. . *It is observed that finding optimal values χ_l and χ_j to maximize the achievable SE is a challenging task due to the non-convex nature of the backhaul capacity. Thus, we develop a low-complexity bandwidth distribution based on the water-filling algorithm (WF-BD) scheme. More specifically, the BD factor can be set as $\gamma_l / \sum_{m=1}^{L} \gamma_m$ for AP l and $\chi_j = \gamma_l / \sum_{m=1}^{J} \gamma_m$ for CPU j, respectively.*

C. Rate-Distortion Theory

To accurately represent an arbitrary real number, an infinite number of bits is required. To this end, representing a continuous random variable with a finite number of bits may introduce distortion, a concept thoroughly analyzed in rate-distortion theory [32]. Let us consider an i.i.d. source $X \sim f_X(x)$, with zero mean and bounded variance $\mathbb{E}\{|X|^2\} = P$. The objective is to compress the entire n-length sequence $\mathbf{X}$ to $\widehat{\mathbf{X}}(m)$, provided that $\mathbb{E}\left\{d(\mathbf{X}, \widehat{\mathbf{X}})\right\} \leq Q$ for a sufficiently large n, where $m \in 1, 2, \ldots, 2^{nC}$ and $d(.,.)$ denotes the distortion measure [32]. Thus, one can perfectly transmit the compression index m through an error-free link with an allocated link rate of C bits/s/Hz. Therefore, based on the [33, Theorem 3.5], the rate distortion function for mmWave fronthaul is as follows:

Theorem 1. *The rate distortion function can be written as*

$$R(Q) = \min_{f(\hat{x}|x):\, \mathbb{E}\left\{|\hat{X}-X|^2\right\} \le Q} I(\hat{X};X). \tag{7}$$

To derive $R(Q)$, it is beneficial to define a test channel $\hat{X} = X + Z_q$, where $Z_q \sim \mathcal{CN}(0,Q)$ is independent of X and characterizes the quantization noise

$$C = I(\hat{X};X) = h(\hat{X}) - h(\hat{X}|X) \overset{(a)}{\le} \log(1+P/Q), \tag{8}$$

where (a) follows from the maximum differential entropy lemma [33, Chapter 2].

Thus, the upper bound (8) overestimates the necessary link rate C required to transfer the quantized signal over the fronthaul/backhaul link. In other words, for a given fronthaul/backhaul rate, we assume the stronger quantization noise with variance $Q^* = \frac{P}{2^C-1}$.

It is worth noting that the correlation among quantization noise components is neglected in the analysis. It has been shown in [34] that such simplification yields negligible error in systems where each AP is equipped with a small number of antennas. Thus, the model in (8) is accurate for the performance analysis considered in this paper.

Remark 2. *Since the proposed Shannon's bound could overestimate the quantization noise, we further explore the scalar quantization to exam the systems performance [35]. More specifically, let C be the scalar quantizer per real sample, the quantization process divides the dynamic range of X into $M = 2^b$ equal-length intervals, each of width Δ. For Δ, we use a common practical assumption that nearly all probability mass of a Gaussian signal is contained within the range $[-\sqrt{P}, \sqrt{P}]$. Thus, the final equation for quantization noise is $Q(b) \approx \frac{\Delta^2}{12} = \frac{1}{12}\cdot\left(\frac{2\sqrt{P}}{2^C}\right)^2 = \frac{P}{3\cdot 2^{2C}}$.*

Remark 3. *Note that analyzing the performance of CF massive MIMO with mmWave fronthaul/backhaul networks differs significantly from that of a wired fronthaul/backhaul network. In wired networks, data are exchanged through lossless links (i.e., ideal channels) with fixed capacities. This means the main controlling factor of the system performance is the capacity of wired links according to which CPU/APs compress the data to be transmitted in the links. In contrast, in mmWave fronthaul/backhaul network, system performance is impacted not only by data compression due to the limited capacity of the links but also by the quality of the mmWave fronthaul/backhaul channels through which the data is transmitted.*

III. Signal Processing Schemes Within CPUs

In the CF massive MIMO with multiple CPUs system, the CPUs can cooperate to process signals to improve system performance. However, a prerequisite for the multiple CPUs cooperation is to transmit the signal from the APs to the associated CPU via the fronthaul links. In this paper, we apply the compress-forward-estimate (CFE) strategies between the APs and the CPUs [22]. More specifically, each AP compresses the received pilot and data signals separately and forwards the compressed versions over the fronthaul link to the associated CPU. The CPU then performs channel estimation and data recoveries. Next, we use CPU j as an example to show the signal processing schemes between the APs and the corresponding CPU.

A. Pilot Transmission and Channel Estimation

For the channel estimation, let $\sqrt{\tau_p}\boldsymbol{\varphi}_k^H \in \mathbb{C}^{\tau_p\times 1}$, where $\|\boldsymbol{\varphi}_k\|^2 = 1$, be the pilot sequence utilized by the UE k. It is considered that $\tau_p < K$. The received signal $\mathbf{y}_{p,l} \in \mathbb{C}^{N\times\tau_p}$ at AP l associated to CPU j is

$$\mathbf{y}_{p,jl} = \sqrt{\tau_p}\sum_{i=1}^{K}\sqrt{p_i}\mathbf{h}_{jli}\boldsymbol{\varphi}_i^T + \mathbf{w}_{p,jl}, \tag{9}$$

where p_i represents the transmitted power of UE i, $\mathbf{w}_{p,j} \in \mathbb{C}^{N\times\tau_p}$ is the receiver noise at AP m with independent $\mathcal{N}_{\mathbb{C}}\left(0,\sigma^2\right)$ entries, and σ^2 denotes the noise power. Then, all AP aspires to transmit these signals to the associated CPUs via fronthaul. CPU j sees

$$\underbrace{\begin{bmatrix}\mathbf{y}_{p,j1}\\ \vdots \\ \mathbf{y}_{p,jL}\end{bmatrix}}_{\triangleq\mathbf{y}_{p,j}} = \sum_{i=1}^{K}\sqrt{p_i}\underbrace{\begin{bmatrix}\mathbf{h}_{j1i}\\ \vdots \\ \mathbf{h}_{jLi}\end{bmatrix}}_{\triangleq\mathbf{h}_{ji}}\varphi_i + \underbrace{\begin{bmatrix}\mathbf{w}_{p,j1}\\ \vdots \\ \mathbf{w}_{p,jL}\end{bmatrix}}_{\triangleq\mathbf{w}_{p,j}} + \underbrace{\begin{bmatrix}\mathbf{q}_{p,j1}\\ \vdots \\ \mathbf{q}_{p,jL}\end{bmatrix}}_{\triangleq\mathbf{q}_{p,j}}, \tag{10}$$

where $\mathbf{w}_{p,j} = [\mathbf{w}_{p,j1},\cdots,\mathbf{w}_{p,jL}]$ and $\mathbf{q}_{fp,j} = [\mathbf{q}_{fp,j1},\cdots,\mathbf{q}_{fp,jL}] \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0},\mathbf{Q}_{fp,j}\right)$ is the quantization noise due to the finite fronthaul capacity. Note that $\mathbf{Q}_{fp,j} = \mathrm{diag}\left(Q_{fp,j1}\mathbf{I}_N,\cdots,Q_{fp,jL}\mathbf{I}_N\right) \in \mathbb{C}^{LN\times LN}$. In addition, $\mathbf{h}_{ji} = [\mathbf{h}_{j1i},\cdots,\mathbf{h}_{jLi}] \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0},\mathbf{R}_{ji}\right)$ is the collective channel between CPU j and UE i, where $\mathbf{R}_{ji} = \mathrm{diag}\left(\mathbf{R}_{j1i},\cdots,\mathbf{R}_{jLi}\right) \in \mathbb{C}^{LN\times LN}$. Based on the above setups, (10) can be rewritten as a more compact form:

$$\mathbf{y}_{p,j} = \sum_{i=1}^{K}\sqrt{p_i}\mathbf{h}_{ji}\boldsymbol{\varphi}_i^T + \mathbf{w}_{p,j} + \mathbf{q}_{fp,j}. \tag{11}$$

To estimate the channel of UE k, CPU j relates the signal $\mathbf{y}_{p,j}$ to the relevant pilot signal $\boldsymbol{\varphi}_k^*$ as

$$\widehat{\mathbf{y}}_{p,j} = \sum_{i=1}^{K}\sqrt{p_i}\mathbf{h}_{ji}\boldsymbol{\varphi}_i^T\boldsymbol{\varphi}_k^* + \mathbf{w}_{p,j}\boldsymbol{\varphi}_k^* + \mathbf{q}_{fp,j}\boldsymbol{\varphi}_k^*. \tag{12}$$

Then, CPU j can estimate the channels by applying minimum mean square error (MMSE) estimator, as

$$\hat{\mathbf{h}}_{jk} = \boldsymbol{\Psi}_{jk}\widehat{\mathbf{y}}_{p,j}, \tag{13}$$

where

$$\boldsymbol{\Psi}_{jk} \triangleq \left(\tau_p\sum_{i=1}^{K}p_i\mathbf{R}_{ji}\left|\boldsymbol{\varphi}_i^T\boldsymbol{\varphi}_k^*\right|^2 + \sigma^2\mathbf{I}_{LN} + \mathbf{Q}_{fp,j}\right)^{-1}\sqrt{p_k\tau_p}\mathbf{R}_{jk}. \tag{14}$$

Note that the channel estimate $\hat{\mathbf{h}}_{jk} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0},\widehat{\mathbf{R}}_{jk}\right)$ with $\widehat{\mathbf{R}}_{jk} = \sqrt{p_k\tau_p}\mathbf{R}_{jk}\boldsymbol{\Psi}_{jk}$ and the channel estimate error $\widetilde{\mathbf{h}}_{jk} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0},\widetilde{\mathbf{R}}_{jk}\right)$, where $\widetilde{\mathbf{R}}_{jk} = \mathbf{R}_{jk} - \widehat{\mathbf{R}}_{jk}$.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3611133

B. Uplink Data Transmission

For the uplink data transmission, the received data signal at AP l in CPU j can be written as

$$\mathbf{y}_{d,jl} = \sum_{i=1}^{K} \mathbf{h}_{jli} s_i + \mathbf{w}_{d,jl}, \tag{15}$$

where $s_i \sim \mathcal{N}_{\mathbb{C}}(0, p_i)$ is the information-bearing signal sent by UE i connected to CPU j with power p_i and $\mathbf{w}_{d,jl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ is the independent noise received at AP l associated to CPU j.

Similar to the channel estimation phase, after receiving the data signal, all APs compress the signals and send to the corresponding CPUs through fronthaul links. The received data signals at CPU j are

$$\mathbf{y}_{d,j} = \sum_{i=1}^{K} \mathbf{h}_{ji} s_i + \mathbf{w}_{d,j} + \mathbf{q}_{fd,j}, \tag{16}$$

where $\mathbf{w}_{d,j} = [\mathbf{w}_{d,j1}, \cdots, \mathbf{w}_{d,jL}]$ and $\mathbf{q}_{fd,j} = [\mathbf{q}_{fd,j1}, \cdots, \mathbf{q}_{fd,jL}] \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{Q}_{fd,j})$ is the quantization noise with $\mathbf{Q}_{fd,j} = (Q_{fd,j1}\mathbf{I}_N, \cdots, Q_{fd,jL}\mathbf{I}_N) \in \mathbb{C}^{LN \times LN}$.

Let $C_{fp,jl}$ and $C_{fd,jl}$ represent the fronthaul rates for forwarding the quantized pilot and data signals from the AP l to the associated CPU j, respectively. It is required that $C_{fp,jl} + C_{fd,jl} = C_{f,jl}$.

Proposition 1. *The fronthaul capacity for pilot transmission is given by*

$$C_{fp,jl} = \frac{\tau_p}{\tau_c} \log_2 \left[1 + \frac{\sum_{i=1}^{K} p_i \mathrm{tr}\left(\mathbf{R}_{jli}\right) + \mathrm{tr}\left(\sigma^2 \mathbf{I}_N\right)}{Q_{fp,jl}} \right], \tag{17}$$

and the fronthaul capacity for uplink data transmission can be written as

$$C_{fd,jl} = \frac{\tau_c - \tau_p}{\tau_c} \log_2 \left[1 + \frac{\sum_{i=1}^{K} p_i \mathrm{tr}\left(\mathbf{R}_{jli}\right) + \mathrm{tr}\left(\sigma^2 \mathbf{I}_N\right)}{Q_{fd,jl}} \right]. \tag{18}$$

Proof: See Appendix A. ∎

Note that Proposition 1 has appeared in previous studies on limited or mmWave fronthaul, e.g., [21], [22], and [30]. We put it here for completeness and to align with the specific assumptions and notations of our system model.

Through this signal processing, each AP needs to send $\tau_p N$ complex scalars for the pilot signals and $(\tau_c - \tau_p)N$ complex scalars for the received signals via fronthaul in each coherence block, resulting in $\tau_c N$ complex scalars in total. The number of complex scalars required for transmission through the fronthaul and backhaul in each coherence block is summarized in Table I.

Remark 4. *After applying the CFE strategies, each CPU has the knowledge of instantaneous CSI for all associated APs. The CPUs can then choose to exchange either instantaneous CSI or statistical CSI with each other for final signal decoding. The specific discussion is presented in the next section.*

TABLE I: Number of complex scalars exchanged from the APs to the corresponding CPU and between the CPUs through the fronthaul and backhaul in each coherence block.

	Fronthaul	Backhaul
Level 4	$\tau_c N$	$((\tau_c - \tau_p)LN + LNK)(J-1)$
Level 3		$(\tau_c - \tau_p)(J-1)K$
Level 2		
Level 1		–

Remark 5. *It is assumed that both the fronthaul and backhaul links operate over the same frequency bands. During system operation, data is first transmitted from the APs to their associated CPUs via the fronthaul, and subsequently exchanged among CPUs through the backhaul. Since these two processes are temporally separated and do not occur simultaneously, it is reasonable to assume that there is no mutual interference between the fronthaul and backhaul links.*

IV. Four Levels Of Multiple CPUs Cooperation

In the multiple CPUs system, all CPUs are interconnected via backhaul links, enabling CPUs to exchange signals and cooperate to improve system performance. Nevertheless, increasing the signal overhead on the backhaul links may result in amplifying the quantization noise from the backhaul. Thus, it is essential to investigate different cooperation schemes among multiple CPUs to provide valuable insights and results for practical CF massive MIMO system with multiple CPUs. To provide a clearer understanding of the proposed four levels of CPU cooperation and the signals exchange in the x-haul links, a comparison of these levels is presented in Fig. 2 (without loss of generality, the CPU performing the final data processing is referred to as "Master CPU") .

As shown in Fig. 2, the data overhead gradually decreases from Level 4 to Level 1, while the degree of centralized coordination increases. Note that the multiple CPUs cooperation schemes is inspired by [10], which investigates the cooperation levels between APs and the CPU. In this paper, we extend the idea to the cooperation among CPUs and further adapt it to the characteristics of mmWave x-haul links.

Notably, the larger data overhead results in more compression noise, whereas higher levels of centralized coordination enhance interference suppression. Therefore, it is crucial to strike a balance between centralized coordination and the signal overhead on the backhaul links.

A. Level 4: Fully Centralized Processing

The highest cooperation level is that all CPUs transmit the signals to one CPU for centralized processing. To calculate the collective combining vector, all CPUs are required to transmit the channel estimation results to Master CPU. The compressed channel estimation at CPU j is as follows:

$$\hat{\mathbf{h}}_{jk}^{'} = \hat{\mathbf{h}}_{jk} + \mathbf{q}_{bp,jk}, \tag{19}$$

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3611133

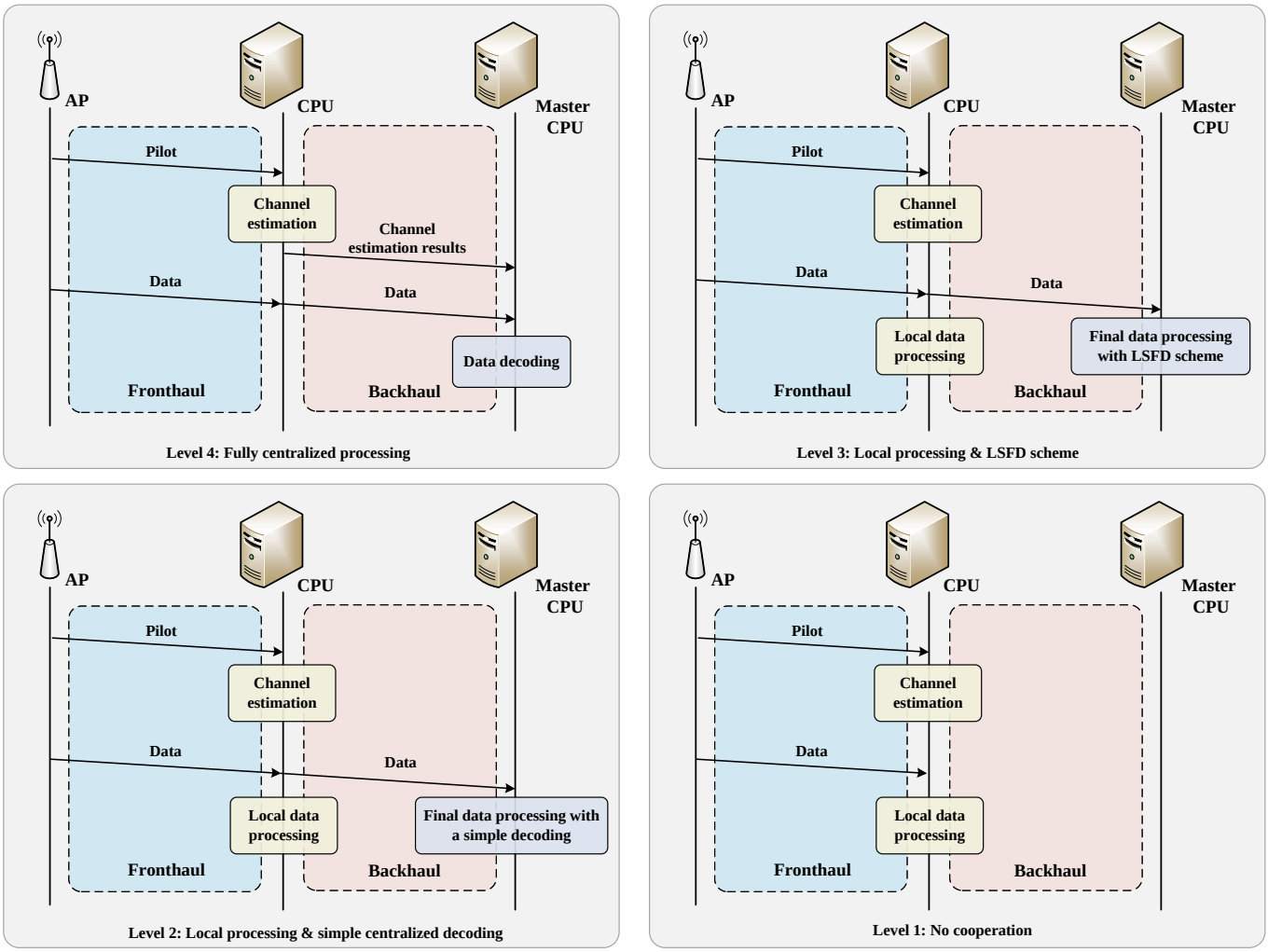


Fig. 2: Four CPUs cooperation levels.

where $\mathbf{q}_{bp,jk} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, Q_{bp,jk}\mathbf{I}_{LN}\right)$ is the quantization noise due to the finite backhaul capacity. Thus, the collective channel estimation $\hat{\mathbf{h}}_k^{'}$ at Master CPU is

$$\underbrace{\begin{bmatrix}\hat{\mathbf{h}}_{1k}^{'}\\ \vdots \\ \hat{\mathbf{h}}_{Jk}^{'}\end{bmatrix}}_{\triangleq \hat{\mathbf{h}}_k^{'}} = \underbrace{\begin{bmatrix}\hat{\mathbf{h}}_{1k}\\ \vdots \\ \hat{\mathbf{h}}_{Jk}\end{bmatrix}}_{\triangleq \hat{\mathbf{h}}_k} + \underbrace{\begin{bmatrix}\mathbf{q}_{bp,1k}\\ \vdots \\ \mathbf{q}_{bp,Jk}\end{bmatrix}}_{\triangleq \mathbf{q}_{bp,k}}, \tag{20}$$

where $\hat{\mathbf{h}}_k^{'} = \left[\hat{\mathbf{h}}_{1k}^{'}, \cdots, \hat{\mathbf{h}}_{Jk}^{'}\right]$ is the collective channel estimation at Master CPU. Note that $\hat{\mathbf{h}}_k^{'} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \widehat{\mathbf{R}}_k^{'}\right)$ with $\widehat{\mathbf{R}}_k^{'} = \mathbb{E}\left\{\hat{\mathbf{h}}_k^{'}\left(\hat{\mathbf{h}}_k^{'}\right)^H\right\}$. Similarly, all the data signals are needed to be transmitted from the other CPUs to Master CPU via backhaul links. The received data signals at Master CPU can be written as

$$\mathbf{y}_d = \sum_{i=1}^{K} \mathbf{h}_i s_i + \mathbf{w}_d + \mathbf{q}_{fd} + \mathbf{q}_{bd}, \tag{21}$$

where $\mathbf{w}_d = [\mathbf{w}_{d,1}, \cdots, \mathbf{w}_{p,J}]$, $\mathbf{q}_{fd} = [\mathbf{q}_{fd,1}, \cdots, \mathbf{q}_{fd,J}] \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{Q}_{fd}\right)$ is the quantization noise due to the finite fronthaul capacity with $\mathbf{Q}_{fd} = \left(\mathbf{Q}_{fd,1}, \cdots, \mathbf{Q}_{fd,J}\right) = \mathrm{diag}\left(Q_{fd,j1}\mathbf{I}_N, \cdots, Q_{fd,jL}\mathbf{I}_N : j = 1, ..., J\right) \in \mathbb{C}^{JLN \times JLN}$, and $\mathbf{q}_{bd} = [\mathbf{q}_{bd,1}, \cdots, \mathbf{q}_{bd,J}] \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{Q}_{bd}\right)$ is the quantization noise due to the finite backhaul capacity with $\mathbf{Q}_{bd} = \left(Q_{bd,1}\mathbf{I}_{LN}, \cdots, Q_{bd,J}\mathbf{I}_{LN}\right) \in \mathbb{C}^{JLN \times JLN}$. Moreover, $\mathbf{h}_i = [\mathbf{h}_{1i}, \cdots, \mathbf{h}_{Ji}] \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}_i\right)$ is the collective channel between Master CPU and UE i, where $\mathbf{R}_i = \mathrm{diag}\left(\mathbf{R}_{1i}, \cdots, \mathbf{R}_{Ji}\right) \in \mathbb{C}^{JLN \times JLN}$. For decoding the final signals, Master CPU can select an arbitrary combining vector $\mathbf{v}_k \in \mathbb{C}^{JLN \times 1}$ for UE k, which is shown as follows:

$$\widehat{s}_k = \mathbf{v}_k^H \mathbf{y}_d = \sum_{i=1}^{K} \mathbf{v}_k^H \mathbf{h}_i s_i + \mathbf{v}_k^H \mathbf{w}_d + \mathbf{v}_k^H \mathbf{q}_{fd} + \mathbf{v}_k^H \mathbf{q}_{bd}. \tag{22}$$

Based on (22), the achievable SE for Level 4 can be obtained by the following proposition:

Proposition 2. *The achievable SE of UE k for Level 4 is*

$$\mathrm{SE}_k^{(4)} = \left(\frac{\tau_c - \tau_p}{\tau_c}\right) \mathbb{E}\left\{\log_2\left(1 + \mathrm{SINR}_k^{(4)}\right)\right\}, \tag{23}$$

$$\mathrm{SINR}_k^{(4)} = \frac{p_k \left| \mathbf{v}_k^H \widehat{\mathbf{h}}_k' \right|^2}{\sum\limits_{i \neq k}^{K} p_i \left| \mathbf{v}_k^H \widehat{\mathbf{h}}_i' \right|^2 + \mathbf{v}_k^H \left(\sum\limits_{i=1}^{K} p_i \widetilde{\mathbf{R}}_i + \mathbf{Q}_{fd} + \mathbf{Q}_{bd} + \sigma^2 \mathbf{I}_{JLN} \right) \mathbf{v}_k} \tag{24}$$

where the instantaneous SINR is presented in (24) *(see top of this page).*

Note that all signals are transmitted to Master CPU through backhaul, thus allowing it to utilizing all CSI to design the combining vector $\mathbf{v}_k$. Two combining schemes are considered for Level 4: MR combining $\mathbf{v}_k = \widehat{\mathbf{h}}_k$ and MMSE combining, which is calculated by minimizing the mean-squared error (MSE) MSE_k with $\mathrm{MSE}_k = \mathbb{E}\left\{ \left| s_k - \mathbf{v}_k^H \mathbf{y}_d \right|^2 \middle| \left\{ \widehat{\mathbf{h}}_i' \right\} \right\}$, as

$$\mathbf{v}_k = p_k \left(\sum_{i=1}^{K} p_i \left(\widehat{\mathbf{h}}_i' \left(\widehat{\mathbf{h}}_i' \right)^H + \widetilde{\mathbf{R}}_i \right) + \mathbf{Q}_{fd} + \mathbf{Q}_{bd} + \sigma^2 \mathbf{I}_{JLN} \right)^{-1} \widehat{\mathbf{h}}_k'. \tag{25}$$

For Level 4, the backhaul capacity $C_{b,j}$ is allocated for pilot transmission $C_{bp,j}$ and dara transmission $C_{bd,j}$. Similar to the fronthaul capacity, it is needed that $C_{b,j} = C_{bp,j} + C_{bd,j}$. The exact value of backhaul capacity is defined as follows:

Proposition 3. *The backhaul capacity of CPU j during the pilot transmission can be written as*

$$C_{bp,j} = \frac{\tau_p}{\tau_c} \log_2 \left[1 + \frac{\sum\limits_{i=1}^{K} \mathrm{tr}\left(\widehat{\mathbf{R}}_{ji} \right)}{Q_{bp,j}} \right], \tag{26}$$

and the backhaul capacity of CPU j during the uplink data transmission is

$$C_{bd,j} = \frac{\tau_c - \tau_p}{\tau_c} \log_2 \left[1 + \frac{\mathbb{E}\left\{ \mathbf{y}_{d,j}^H \mathbf{y}_{d,j} \right\}}{Q_{bd,j}} \right]. \tag{27}$$

In each coherence block, all CPUs except Master CPU need to transmit $(\tau_c - \tau_p)LN$ complex scalars for data transmission and LNK for channel estimation results transmission, becoming $((\tau_c - \tau_p)LN + LNK)(J-1)$ in total, which is shown in Table I.

Remark 6. *Note that Master CPU does not need to transmit information to itself. Therefore, the backhaul capacity of Master CPU $C_{b,MC}$ can be seen as infinite, and equivalently the corresponding quantization noise $Q_{b,MC}$ given in* (21) *tends to zero, i.e., $C_{b,MC} = \infty$ and $Q_{b,MC} = 0$.*

Remark 7. *It can be observed that the quantization noise due to the limited fronthaul link is amplified during the signal compression processing in the backhaul link. This means that backhaul links exacerbate the impact of finite fronthaul, which can seriously affect the system performance.*

B. Level 3: Local Processing & Large-Scale Fading Decoding

Although Level 4 can suppress the interference with collective combining at Master CPU, it assigns all computational tasks to Master CPU. This may place excessive demands on Master CPU while wasting the processing capabilities of the other CPUs. Thus, we consider each CPU can preprocess its signal and Master CPU only needs to do the final decoding. More specifically, in Level 3, each CPU processes the signals using the local CSI, and then transmits the signals via backhaul to Master CPU for final decoding. The local estimate of the signal s_k at CPU j is

$$\widehat{s}_{jk} = \mathbf{v}_{jk}^H \mathbf{y}_{d,j} = \sum_{i=1}^{K} \mathbf{v}_{jk}^H \mathbf{h}_{ji} s_i + \mathbf{v}_{jk}^H \mathbf{w}_{d,j} + \mathbf{v}_{jk}^H \mathbf{q}_{fd,j}, \tag{28}$$

where $\mathbf{v}_{jk} \in \mathbb{C}^{LN \times 1}$ is the local combining vector calculated by CPU j. Note that arbitrary combining vector can be used in the above equation. Nevertheless, only local estimation $\{\mathbf{h}_{ji} : i = 1, ..., K\}$ is available in CPU j. Similar to Level 4, two local combining schemes are considered: the simple MR combining $\mathbf{v}_{jk} = \widehat{\mathbf{h}}_{jk}$ and the local MMSE (L-MMSE) combining, which is calculated by minimizing the MSE, $\mathrm{MSE}_{jk} = \mathbb{E}\{|s_k - \mathbf{v}_{jk}^H \mathbf{y}_{d,j}|^2 \big| \{\hat{\mathbf{h}}_{ji}\}\}$, is

$$\mathbf{v}_{jk} = p_k \left(\sum_{i=1}^{K} p_i \left(\widehat{\mathbf{h}}_{ji} \widehat{\mathbf{h}}_{ji}^H + \widetilde{\mathbf{R}}_{ji} \right) + \mathbf{Q}_{fd,j} + \sigma^2 \mathbf{I}_{LN} \right)^{-1} \widehat{\mathbf{h}}_{jk}. \tag{29}$$

A two-layer decoding scheme is considered at Level 3. After the first layer decoding, all CPUs transmit the decoded signals to Master CPU for the second layer decoding through the backhaul links. Let $\left\{ a_{jk}^* : j = 1, ..., J \right\}$ represent the linearly combined weights, the received signals at Master CPU is given by

$$\begin{aligned} \hat{s}_k = \sum_{j=1}^{J} a_{jk}^* \hat{s}_{jk} = & \sum_{i=1}^{K} \sum_{j=1}^{J} a_{jk}^* \mathbf{v}_{jk}^H \mathbf{h}_{ji} s_i + \sum_{j=1}^{J} a_{jk}^* \mathbf{v}_{jk}^H \mathbf{w}_{d,j} \\ & + \sum_{j=1}^{J} a_{jk}^* \mathbf{v}_{jk}^H \mathbf{q}_{fd,j} + \sum_{j=1}^{J} a_{jk}^* q_{bd,j}, \end{aligned} \tag{30}$$

where $q_{bd,j} \sim \mathcal{CN}(0, Q_{bd,j})$ is the quantization noise of CPU j due to the backhaul link. Then we can rewrite (30) in a more compact form as

$$\hat{s}_k = \mathbf{a}_k^H \mathbf{g}_{kk} s_k + \sum_{i \neq k}^{K} \mathbf{a}_k^H \mathbf{g}_{ki} s_i + \mathbf{a}_k^H \mathbf{w}_{d,k} + \mathbf{a}_k^H \mathbf{q}_{fd,k} + \mathbf{a}_k^H \mathbf{q}_{bd,k}, \tag{31}$$

where $\mathbf{a}_k = [a_{1k} \dots a_{Jk}]^T \in \mathbb{C}^{J \times 1}$ is the weighting coefficient vector, $\mathbf{w}_{d,k} = [\mathbf{v}_{1k}^H \mathbf{w}_{d,1} \dots \mathbf{v}_{Jk}^H \mathbf{w}_{d,J}]^T \in \mathbb{C}^{J \times 1}$, $\mathbf{q}_{fd,k} = [\mathbf{v}_{1k}^H \mathbf{q}_{fd,1} \dots \mathbf{v}_{Jk}^H \mathbf{q}_{fd,J}]^T \in \mathbb{C}^{J \times 1}$, and $\mathbf{q}_{bd,k} = [\mathbf{v}_{1k}^H \mathbf{q}_{bd,1} \dots \mathbf{v}_{Jk}^H \mathbf{q}_{bd,J}]^T \in \mathbb{C}^{J \times 1}$. Furthermore, $\mathbf{g}_{ki} = [\mathbf{v}_{1k}^H \mathbf{h}_{1i} \dots \mathbf{v}_{Jk}^H \mathbf{h}_{Ji}]^T$ characterizes the L-dimensional vector for the receive-combined channels between UE k and each of the CPUs. Then, from (31), the achievable SE for Level 3 can

$$\mathrm{SINR}_k^{(3)} = \frac{p_k \left| \mathbf{a}_k^H \mathbb{E}\left\{\mathbf{g}_{kk}\right\} \right|^2}{\sum\limits_{i=1}^{K} p_i \mathbb{E}\left\{ \left| \mathbf{a}_k^H \mathbf{g}_{ki} \right|^2 \right\} - p_k \left| \mathbf{a}_k^H \mathbb{E}\left\{\mathbf{g}_{kk}\right\} \right|^2 + \mathbb{E}\left\{ \left| \mathbf{a}_k^H \mathbf{w}_{d,k} \right|^2 \right\} + \mathbb{E}\left\{ \left| \mathbf{a}_k^H \mathbf{q}_{fd,k} \right|^2 \right\} + \mathbb{E}\left\{ \left| \mathbf{a}_k^H \mathbf{q}_{bd,k} \right|^2 \right\}} \tag{33}$$

be computed as follows:

Proposition 4. *An achievable SE of UE k at Level 3 is given by*

$$\mathrm{SE}_k^{(3)} = \left(\frac{\tau_c - \tau_p}{\tau_c} \right) \log_2 \left(1 + \mathrm{SINR}_k^{(3)} \right), \tag{32}$$

where the effective SINR is shown in (33) *(see top of this page).*

A second layer decoding structure, which can be called "LSFD", is investigated at Level 3. Note that only statistical CSIs are available at CPU, we utilize the use-and-then-forget (UatF) bound to calculate the achievable SE for the LSFD scheme as follows:

Corollary 1. *The deterministic weighting vector $\mathbf{a}_k$ for maximizing the achievable SE is*

$$\begin{aligned} \mathbf{a}_k = & \Big(\sum_{i=1}^{K} p_i \mathbb{E}\left\{ \mathbf{g}_{ki} \mathbf{g}_{ki}^H \right\} - p_k \mathbb{E}\left\{ \mathbf{g}_{kk} \right\} \mathbb{E}\left\{ \mathbf{g}_{kk}^H \right\} \\ & + \mathbb{E}\left\{ \mathbf{w}_{d,k} \right\} \mathbb{E}\left\{ \mathbf{w}_{d,k}^H \right\} + \mathbb{E}\left\{ \mathbf{q}_{fd,k} \right\} \mathbb{E}\left\{ \mathbf{q}_{fd,k}^H \right\} \\ & + \mathbb{E}\left\{ \mathbf{q}_{bd,k} \right\} \mathbb{E}\left\{ \mathbf{q}_{bd,k}^H \right\} \Big)^{-1} \mathbb{E}\left\{ \mathbf{g}_{kk} \right\}, \end{aligned} \tag{34}$$

which results in the maximum value as

$$\begin{aligned} \mathrm{SINR}_k^{(3)} = & p_k \mathbb{E}\left\{ \mathbf{g}_{kk}^H \right\} \Big(\sum_{i=1}^{K} p_i \mathbb{E}\left\{ \mathbf{g}_{ki} \mathbf{g}_{ki}^H \right\} - p_k \mathbb{E}\left\{ \mathbf{g}_{kk} \right\} \mathbb{E}\left\{ \mathbf{g}_{kk}^H \right\} \\ & + \mathbb{E}\left\{ \mathbf{w}_{d,k} \right\} \mathbb{E}\left\{ \mathbf{w}_{d,k}^H \right\} + \mathbb{E}\left\{ \mathbf{q}_{fd,k} \right\} \mathbb{E}\left\{ \mathbf{q}_{fd,k}^H \right\} \\ & + \mathbb{E}\left\{ \mathbf{q}_{bd,k} \right\} \mathbb{E}\left\{ \mathbf{q}_{bd,k}^H \right\} \Big)^{-1} \mathbb{E}\left\{ \mathbf{g}_{kk} \right\}. \end{aligned} \tag{35}$$

Note that it follows from [2, Lemma B.10] by the observation that (33) *is a generalized Rayleigh quotient with respect to the weighting vector $\mathbf{a}_k$.*

For Level 3, the backhaul links is only used for the data transmission, i.e., $C_{bd,j} = C_{b,j}$. Thus, the backhaul capacity of CPU j is defined as

Proposition 5. *The backhaul capacity of CPU j at Level 3 is*

$$C_{bd,j} = \log_2 \left[1 + \frac{\mathbb{E}\left\{ \left| \hat{s}_{jk} \right|^2 \right\}}{Q_{bd,j}} \right]. \tag{36}$$

It is worth noting that the closed-form expressions of SINR given in (33) cannot be derived applying the L-MMSE combining, due to the presence of random matrices in the inverse matrix. Nevertheless, if the MR combining $\mathbf{v}_{jk} = \widehat{\mathbf{h}}_{jk}$ is used, we can calculate the expectations given in (33) over closed-form and derive the closed-form SE expression as follows:

Theorem 2. *The closed-form expression for the SINR given in* (33) *can be written as*

$$\sum_{j=1}^{J} \mathbb{E}\left\{ \sqrt{p_k} \widehat{\mathbf{h}}_{jk}^H \mathbf{h}_{jk} \right\} = \sqrt{p_k} \sum_{j=1}^{J} \mathrm{tr}\left(\widehat{\mathbf{R}}_{jk} \right), \tag{37}$$

$$\begin{aligned} \mathbb{E}\left\{ \left| \sum_{j=1}^{J} \sqrt{p_i} \widehat{\mathbf{h}}_{jk}^H \mathbf{h}_{ji} \right|^2 \right\} = & p_i \sum_{j=1}^{J} \mathrm{tr}\left(\widehat{\mathbf{R}}_{jk} \mathbf{R}_{ji} \right) \\ & + p_i \left| \boldsymbol{\varphi}_i^T \boldsymbol{\varphi}_k^* \right|^2 \left| \sum_{j=1}^{J} \mathrm{tr}\left(\frac{\widehat{\mathbf{R}}_{jk}}{\mathbf{R}_{jk}} \mathbf{R}_{ji} \right) \right|^2, \end{aligned} \tag{38}$$

$$\mathbb{E}\left\{ \left| \sum_{j=1}^{J} \widehat{\mathbf{h}}_{jk}^H \mathbf{w}_{d,j} \right|^2 \right\} = \sum_{j=1}^{J} \mathrm{tr}\left(\widehat{\mathbf{R}}_{jk} \sigma^2 \right), \tag{39}$$

$$\mathbb{E}\left\{ \left| \sum_{j=1}^{J} \widehat{\mathbf{h}}_{jk}^H \mathbf{q}_{fd,j} \right|^2 \right\} = \sum_{j=1}^{J} \mathrm{tr}\left(\widehat{\mathbf{R}}_{jk} \mathbf{Q}_{fd,j} \right), \tag{40}$$

and

$$\mathbb{E}\left\{ \left| \sum_{j=1}^{J} q_{bd,j} \right|^2 \right\} = \sum_{j=1}^{J} Q_{fd,j}. \tag{41}$$

Proof. See Appendix B. □

C. Level 2: Local Processing & Simple Centralized Decoding

Although the SE can be maximized by using LSFD at Level 3, it requires the knowledge of a number of statistical parameters, which could be very large in CF massive MIMO. As an alternative, we can simply take the average of the local estimates. Thus, the local estimate at Master CPU is

$$\hat{s}_k = \sum_{j=1}^{J} \frac{1}{J} \hat{s}_{jk}, \tag{42}$$

where $\hat{s}_{jk}$ is given in (28) and can be achieved by any local combining vector. Note that it is equivalent to set the weighting coefficient vector $\mathbf{a}_k$ as $\mathbf{a}_k = [1/J \ldots 1/J]^T$, the result of SE is as follows:

Corollary 2. *An achievable SE of UE k at Level 2 can be written as*

$$\mathrm{SE}_k^{(2)} = \left(\frac{\tau_c - \tau_p}{\tau_c} \right) \log_2 \left(1 + \mathrm{SINR}_k^{(2)} \right), \tag{43}$$

where the effective SINR is given in (44) *(see top of the next page).*

It is worth noting that the closed-form SE expressions of Level 2 can also be derived using the MR combining.

$$\mathrm{SINR}_k^{(2)} = \frac{p_k \left| \sum_{j=1}^{J} \mathbb{E}\left\{ \mathbf{v}_{jk}^H \mathbf{h}_{jk} \right\} \right|^2}{\sum_{i=1}^{K} p_i \mathbb{E}\left\{ \left| \sum_{j=1}^{J} \mathbf{v}_{jk}^H \mathbf{h}_{ji} \right|^2 \right\} - p_k \left| \sum_{j=1}^{J} \mathbf{v}_{jk}^H \mathbf{h}_{jk} \right|^2 + \mathbb{E}\left\{ \left| \sum_{j=1}^{J} \mathbf{v}_{jk}^H \mathbf{w}_{d,j} \right|^2 \right\} + \mathbb{E}\left\{ \left| \sum_{j=1}^{J} \mathbf{v}_{jk}^H \mathbf{q}_{fd,j} \right|^2 \right\} + \mathbb{E}\left\{ \left| \sum_{j=1}^{J} \mathbf{q}_{bd,j} \right|^2 \right\}} \tag{44}$$

Nevertheless, we omit it because it is similar to the closed-form expression of Level 3.

For Level 2 and Level 3, all CPUs except Master CPU are required to send $(\tau_c - \tau_p)K$ complex scalars in each coherence block, which become $(\tau_c - \tau_p)(J-1)K$ in total. This result is illustrated in Table I.

Remark 8. *From* (33) *and* (44)*, it is evident that the combining vector is not correlated with the associated quantization noise introduced by the backhaul link at both Level 2 and Level 3. As a result, the combining process is unable to effectively suppress the quantization noise caused by the backhaul link, leading to the SE loss.*

D. Level 1: No Cooperation

The lowest level of cooperation is no signal exchange on the backhaul link, resulting in all CPUs having to decode the signals alone. In this case, the decoding is done locally at the CPU by using the locally instantaneous CSI, potentially making it still competitive. Based on the equation given in (28), the achievable SE at Level 1 is shown as follows:

Corollary 3. *An achievable SE of UE k can be written as*

$$\mathrm{SE}_k^{(1)} = \left(1 - \frac{\tau_p}{\tau_c}\right) \max_{j \in \{1,\ldots,J\}} \mathbb{E}\left\{ \log_2\left(1 + \mathrm{SINR}_{jk}^{(1)}\right) \right\}, \tag{45}$$

where the instantaneous effective SINR of CPU j is given by

$$\mathrm{SINR}_{jk}^{(1)} = \frac{p_k |\mathbf{v}_{jk}^H \widehat{\mathbf{h}}_{jk}|^2}{\sum_{i \neq k}^{K} p_i |\mathbf{v}_{jk}^H \widehat{\mathbf{h}}_{ji}|^2 + \mathbf{v}_{jk}^H (\sum_{i=1}^{K} p_i \mathbf{R}_{ji} + \mathbf{Q}_{fd,j} + \sigma^2 \mathbf{I}_{LN}) \mathbf{v}_{jk}}. \tag{46}$$

Since no signal exchange occurs on the backhaul link in this level, no compression noise is caused by the backhaul link. Note that Level 1 is applied in [30], and it is compared with the other levels in the section VI.

V. Backhaul Capacity Allocation

For Level 4, the backhaul capacity $C_{b,j}$ is allocated for pilot transmission $C_{bp,j}$ and data transmission $C_{bd,j}$, i.e., $C_{b,j} = C_{bp,j} + C_{bd,j}$. To construct the optimization problem, we introduce the backhaul capacity allocation (BCA) factor $\lambda \in (0,1)$, resulting in the following equation:

$$\begin{cases} C_{b,j} = \lambda C_{bd,j} \\ C_{b,j} = (1-\lambda) C_{bp,j} \end{cases}. \tag{47}$$

Thus, the optimization problem can be formulated as

$$\max_{\{\lambda\}} \mathrm{SSE}^{(4)} = \sum_{i=1}^{K} \mathrm{SE}_i^{(4)} \quad \text{subject to } 0 \leq \lambda \leq 1. \tag{48}$$

It is found that as BCA factor λ increases, the compression noise during the channel estimation phase increases, while that during the data transmission phase undergoes a decrease. Thus, it is essential to find an appropriate factor λ for reducing the impact of limited backhaul link[1]. To this end, we propose a bisection search for solving (48), in which a sequence of convex feasibility problems is resolved in each step.

Binary search is an algorithm used to find the position of a target value within a sorted array. The algorithm works by comparing the target value with the middle element of the array. If they are not equal, the half of the array where the target value cannot be eliminated. The search then continues on the remaining half, comparing the target value to the new middle element. This process is repeated until the target value is located. By consistently discarding the half where the target value cannot be, the algorithm efficiently narrows down the search area with each iteration [36].

In this section, we use the binary search algorithm for solving (48). Initially, the algorithm sets $\lambda_{\min} = 0$ and $\lambda_{\max} = 1$ along with a tolerance ε. It computes the sum SEs at the boundaries and selects the maximum as the initial value. Then, it is needed to iteratively narrow the search range by updating $\lambda_{\min}$ and $\lambda_{\max}$ based on the sum SEs of the midpoint and a slightly increased value. The algorithm continues until the difference between $\lambda_{\max}$ and $\lambda_{\min}$ is smaller than the tolerance ε. The final value of λ is the result that maximizes the sum SEs. The specific steps for solving (48) are outlined in Algorithm 1. In addition, it is worth noting that the fronthaul capacity is required to be allocated for pilot transmission and data transmission at all levels, the proposed binary search algorithm is also applied for finding the valuable fronthaul capacity allocation factor.

VI. Numerical Results

A. Simulation Setup and Radio Propagation Model

We consider that J CPUs and K UEs are located in the area, in which the distance between the other CPUs and Master CPU is 500 m. Moreover, L APs equipped with N antennas are uniformly distributed at the area of a circle centered at them corresponding CPU with a radius of 500 m. Note that the total number of APs in this area is $J \times L$.

[1]Note that the backhaul capacity is only allocated for data transmission at Level 2 and Level 3, so is no necessity for capacity allocation at these levels.

Algorithm 1 Proposed Binary Search Algorithm for Solving (48)

Result: the valuable BCA factor λ.

Input: $\text{SSE}^{(4)}$, the tolerance ε, the minimum value λ_{min}, and the maximum value λ_{max}.

Initialization: Set the initial values $\lambda_{\text{min}} = 0$ and $\lambda_{\text{max}} = 1$; Set the tolerance $\varepsilon > 0$ and an increase $0 < \Delta\lambda \ll \varepsilon$; $\text{SSE}_{\text{max}}^{(4)} = \text{SSE}^{(4)}(\lambda^*) = \max\left\{\left[\text{SSE}^{(4)}(\lambda_{\text{min}}), \text{SSE}^{(4)}(\lambda_{\text{max}})\right]\right\}$ and $\lambda = \lambda^*$;

while $\lambda_{\text{max}} - \lambda_{\text{min}} > \varepsilon$ **do**

Set $\lambda_{\text{next}} = (\lambda_{\text{max}} + \lambda_{\text{min}})/2$;

$\text{SSE}_{\text{next}}^{(4)} = \text{SSE}^{(4)}(\lambda_{\text{next}})$ and $\text{SSE}_{\Delta}^{(4)} = \text{SSE}^{(4)}(\lambda_{\text{next}} + \Delta\lambda)$;

If $\text{SSE}_{\Delta}^{(4)} > \text{SSE}_{\text{next}}^{(4)}$, then set $\lambda_{\text{min}} = \lambda_{\text{next}}$, else set $\lambda_{\text{max}} = \lambda_{\text{next}}$;

If $\text{SSE}_{\text{next}}^{(4)} > \text{SSE}_{\text{max}}^{(4)}$, then set $\text{SSE}_{\text{max}}^{(4)} = \text{SSE}_{\text{next}}^{(4)}$ and $\lambda = \lambda_{\text{next}}$;

end while

Output: λ.

For large-scale fading, the classical 3GPP Urban Microcell model [10] with a 2 GHz carrier frequency is used as

$$\beta_{xy}[\text{dB}] = -30.5 - 36.7\log_{10}\left(\frac{d_{xy}}{1\text{ m}}\right) + Q_{xy}, \tag{49}$$

where d_{xy} is the 3D distance between AP x and UE y, which accounts for the border wraparound and AP antenna mounted at a height of 10 m. Moreover, $Q_{xy} \sim \mathcal{CN}\left(0, 4^2\right)$ denotes the shadow fading, and the shadowing terms between AP x and different UEs are stipulated by

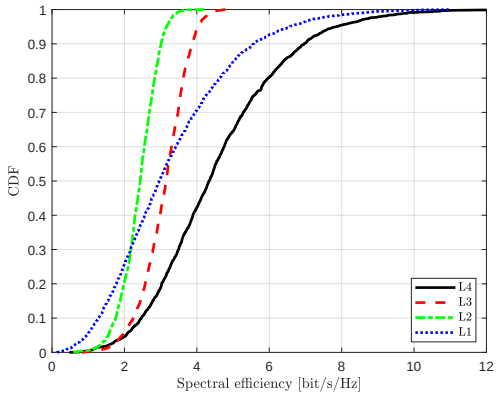
$$\mathbb{E}\left\{Q_{kx}Q_{in}\right\} = \begin{cases} 4^2 2^{-\delta_{ki}/9\text{ m}}, & l = n \\ 0, & l \neq n \end{cases}, \tag{50}$$

where δ_{ki} is the distance between UE k and UE i. The second term in (50) characterizes the correlation of shadowing terms associated with two distinct APs. This correlation is negligible due to the simulation setup, where there is a minimum separation of at least 50 m between adjacent APs.

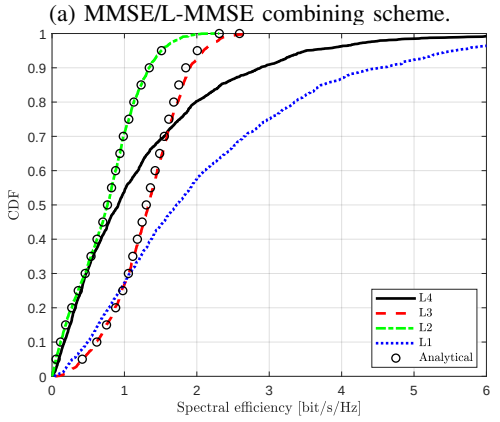
Furthermore, for the backhaul/fronthaul network communication under mmWave band, the transmission noise variance in the backhaul/fronthaul is $\sigma_w^2 = 290 \times \kappa \times \mathcal{B}_A \times F$, where $\kappa = 1.380649 \times 10^{-23}$ is the Boltzman constant. The system bandwidth is $\mathcal{B}_A = 20$ MHz, and $F = 8.7$ dB is the noise figure [35]. Furthermore, we assume a path-loss exponent of $\alpha = 2.92$ and $A = 72$ dB with available contiguous bandwidth $\mathcal{B}_f$ and $\mathcal{B}_b$ up to 28 GHz as in [31]. The total antenna gain for the backhaul/fronthaul link is set to $G_t = 10$ dB, and $P_f = 1$ W is the transmit power [30]. All simulation parameters are shown in Table II.

TABLE II: Simulation Parameters

Parameter	Value
Communication bandwidth, $\mathcal{B}_A$	20 MHz
Available mmWave bandwidth, $\mathcal{B}_f$ and $\mathcal{B}_b$	28 GHz
Noise figure, F	8.7 dB
Path-loss intercept, A	72 dB
Path-loss exponent, α	2.92
AP antenna height	10 m
Noise power for communication, σ^2	-94 dBm
Coherence time, τ_c	200 msec
Uplink training duration, τ_p	6 msec
Number of CPUs, J	5
Number of APs per CPU, L	12
Number of antennas per AP, N	2
Number of UEs, K	12
Uplink transmit power per UE, p_k	20 dBm



(a) MMSE/L-MMSE combining scheme.



(b) MR combining scheme.

Fig. 3: CDF of SE per UE for MMSE/L-MMSE combining and MR combining with J=5, L=12, N=2, and $K = 12$.

B. SE of Different Multiple CPUs Cooperation

We firstly investigate the function of multiple CPUs cooperation. Fig. 3 presents the cumulative distribution function (CDF) of the uplink SE per UE for four CPU cooperation levels over MMSE/L-MMSE combining and MR combining with limited fronthaul and backhaul when J=5, L=12, N=2, and $K = 12$. From Fig. 3(a), it is evident that even though more compression noise is generated, Level 4 still achieves the highest SE with MMSE combining. Since the compression noise can be suppressed by LSFD, the SE at Level 3 is second only to it at Level 4. Moreover, as there is no compression noise from backhaul link at Level 1, it is useful when UEs have good channel conditions. This is due to the fact that the curve of Level 1 crosses the curve of Level 2 at 30% likely SE.

Fig. 3(b) examines the performance of the proposed system

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3611133

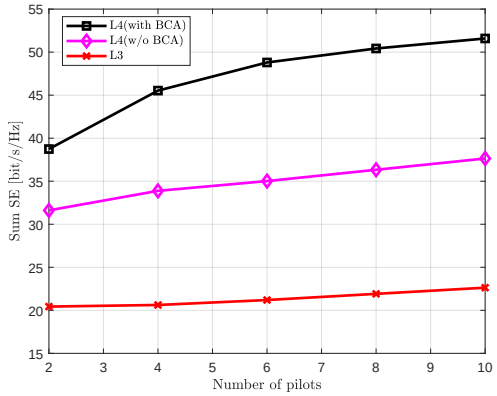


Fig. 4: Sum SE with MMSE/L-MMSE combining for Level 3 and Level 4 with $K = 12$ and $\tau_c = 200$.

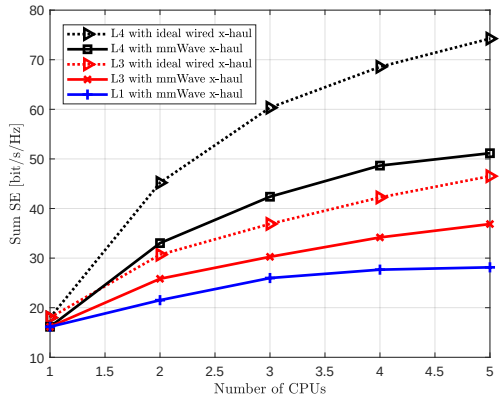


Fig. 5: Sum SE under different numbers of CPUs with $K = 12$ and $L = 12$.

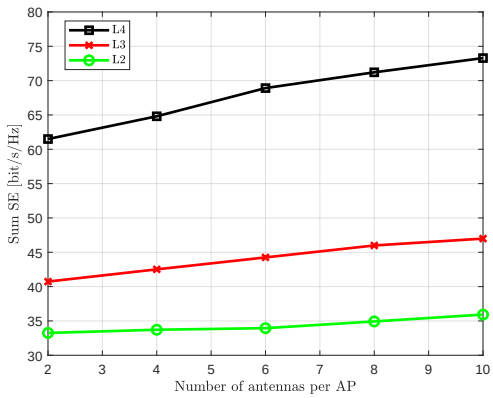


Fig. 6: Sum SE under different number of antennas per AP with $L = 12$.

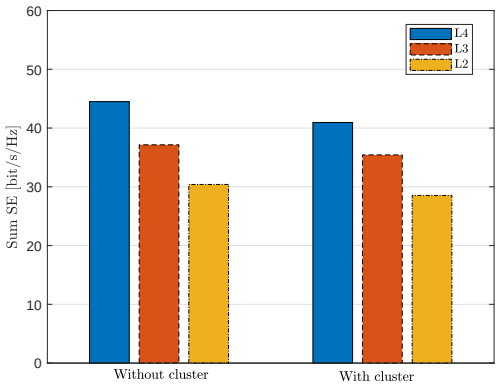


Fig. 7: Sum SE under different multiple CPUs cooperation schemes with $J = 5$.

when MR combining scheme is applied. Due to the weakened immunity to compression noise and interference, the SEs decrease at all levels. Particularly, Level 4 is almost coincident with Level 2 at 95% likely SE since a large amounts of compression noise cannot be eliminated. Moreover, Level 1 shows the sub-optimal performance due to the lowest amount of compression noise created. In addition, Level 3 remains competitive in mitigating compression noise and interference through the usage of LSFD.

Fig. 4 presents the sum SE with MMSE/L-MMSE combining of Level 3 and Level 4, considering the usage of BCA for Level 4. It can be found that Level 4 with BCA can outperform the Level 3 and Level 4 without BCA. We also find that the advantage of it can be further amplified when the number of pilots is sufficient. Another observation from Fig. 4 is that both Level 3 and Level 4 can benefit from the increase in the number of pilots, proving the importance of valuable combining schemes.

Fig. 5 shows the sum SE with MMSE/L-MMSE combining under different numbers of CPUs. It can be observed that as the number of CPUs increases, the SE improves across all levels, including Level 1. This is because Level 1 can select a better CPU for service. However, the upward trend quickly plateaus as an adequate number of CPUs becomes available for selection. Meanwhile, the SE of Level 3 and Level 4 continues to rise, highlighting the importance of multiple CPUs cooperation. In addition, it is observed that the performance gap in SEs between ideal wired and mmWave x-haul links is larger at Level 4 than at Level 3. This is because Level 4 introduces more quantization noise, making the impact of increasing the number of CPUs more significant.

Fig. 6 illustrates the sum SE for different numbers of antennas per AP using MMSE/L-MMSE combining. It is observed that the SE increases across all levels as the number of antennas per AP grows, indicating that the proposed schemes remain effective when the number of antennas per AP is large.

Fig. 7 presents the sum SE of CF massive MIMO under different multiple CPUs cooperation schemes. We take the user-centric concept into consideration and achieve it through dynamic cooperation clustering [9]. It can be found that dynamic cooperation clustering can obtain about 90% SE compared to the system without cluster, which shows that it has a certain positive effect on CF massive MIMO systems when x-haul links is connected by mmWave. This is due to the fact that a small cluster size aids in reducing the compression noise introduced bu the distant APs or CPUs. Therefore, dynamic cooperation clustering can be adopted to assist multiple CPUs cooperation schemes for enhancing the scalability of systems.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3611133

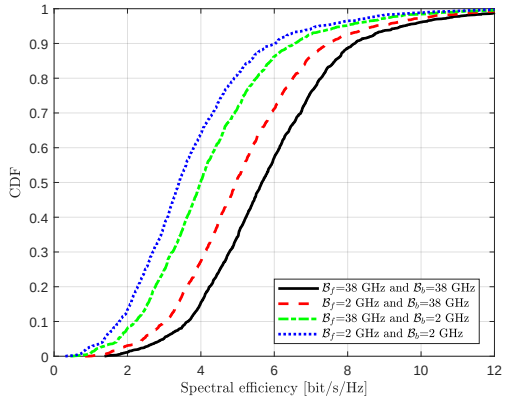


Fig. 8: CDF of SE per UE with MMSE/L-MMSE combining with $K = 12$, $L = 12$, and $J = 5$.

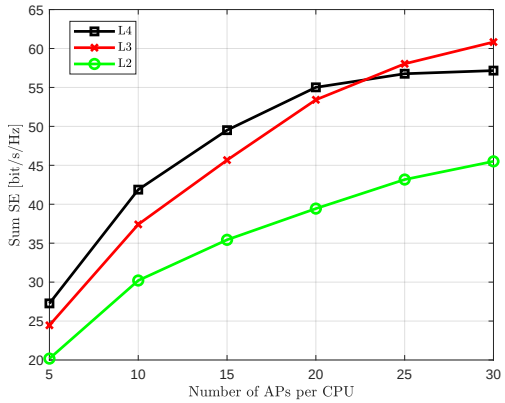


Fig. 9: Sum SE under varying numbers of APs connected to the CPUs with $K = 12$ and $J = 5$.

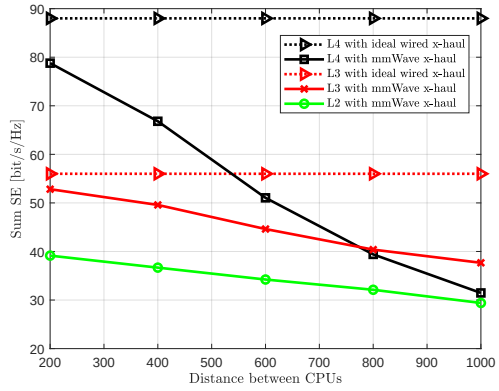


Fig. 10: Sum SE under different distance between CPUs with $J = 5$.

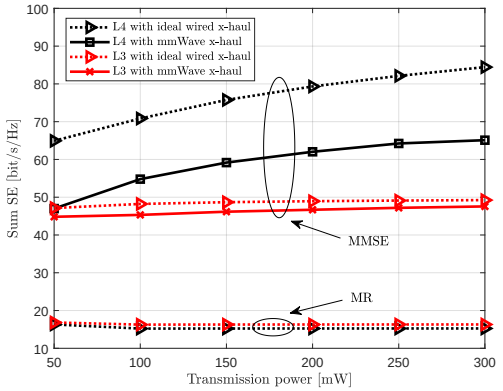


Fig. 11: Sum SE under different number of antennas per AP with $L = 12$.

C. Effects of MmWave Fronthaul and Backhaul

Based on the above simulation results, we have found that it is important to exchange the signals via the backhaul link for multiple CPUs cooperation. However, this can also result in additional compression noise due to the consideration of mmWave backhaul link. Thus, we provide some insights and results into the effect of mmWave fronthaul and backhaul.

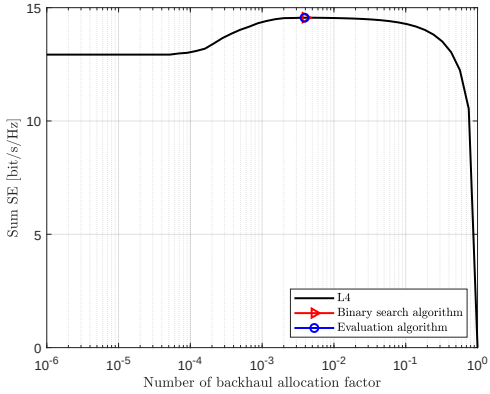
Fig. 8 illustrates the CDF of SE per UE for Level 4 under different bandwidth of fronthaul and backhaul. We observe that reducing the backhaul link bandwidth results in a more significant SE loss. This indicates that limited backhaul has a greater impact on system performance compared to limited fronthaul.

Fig. 9 depicts the sum SE with MMSE/L-MMSE combining under varying number of APs connected to the CPUs. Along with the increase in the total number of APs comes a corresponding increase in the compression noise. It is clear that when the number of APs is high, at Level 4, the increase in the number of APs cannot compensate for the SE loss caused by the additional compression noise. As a result, the SE at Level 4 cannot be further enhanced and thus be underperformed by the SE at Level 3.

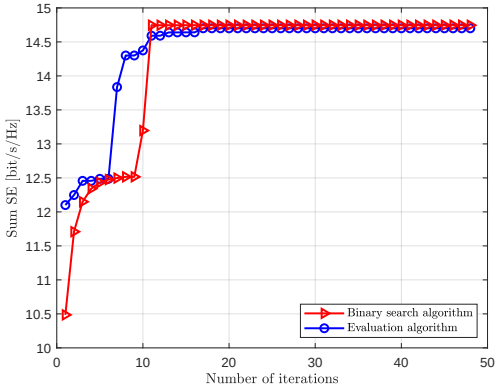
It is worth noting that increasing the distance between CPUs can reduce backhaul capacity, thereby affecting the SE of each cooperation level. Fig. 10 shows the sum SE with MMSE/L-MMSE combining under varying distances between CPUs. Note that systems with ideal wired backhaul links are not affected by the distance between CPUs, as no compression noise is introduced. Simulation results show that the SEs of all levels with mmWave backhaul decrease due to increased compression noise. Moreover, Level 4 declines significantly more than other levels, revealing that it is more affected by mmWave backhaul. In contrast, Level 3 shows the slowest decrease, as its LSFD effectively suppresses interference. In addition, when the distance is large, Level 3 can outperform Level 4. This underscores the importance of carefully considering signal processing schemes in CF massive MIMO systems with multiple CPUs, particularly when the distance between CPUs is large.

Fig. 11 shows the sum SE of system with respect to transmit power, ranging from 50 to 300 mW. It is clear that Level 4 can effectively boost communication quality by increasing the transmission power. But Level 3 cannot further enhance the SE, implying that transmission power is not the limiting factor for Level 3. Moreover, The usage of the mmWave x-haul links does indeed result in the performance losses. Nevertheless, it is still possible to make up for the loss by using appropriate approachs, e.g., the MMSE combining with mmWave x-haul can consistently outperform the MR combining with ideal wired x-haul.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3611133



(a) Sum SE versus allocated capacity.



(b) Number of iterations for convergence with MMSE combining.

Fig. 12: Comparisons of the proposed binary search algorithm with the evaluation algorithm.

Fig. 12(a) depicts the sum SE versus allocated capacity. Note that when the capacity allocation factor $\lambda \to 1$, all capacities are required for the pilot transmission, leaving little data can be transmitted, i.e., sum SE tends to 0. When $\lambda \to 0$, data can be transmitted unhindered but the channel estimation is almost unusable. In this case, the beamforming vectors can be regarded as random. Nevertheless, a few "fortunate" UEs may still succeed in transmitting their signals, and therefore, the sum SE does not drop to zero. Furthermore, it is clear that there exists an optimal value, and both the proposed binary search and evaluation algorithms can approximate this value. Fig.12(b) shows the number of iterations for convergence with MMSE combining. It is found that the proposed binary search algorithm is able to converge faster with the similar performance.

Fig. 13 compares the average SE of Level 3 and Level 4 with and without WF-BD scheme across varying UEs. It is observed that WF-BD can obtain about 12% and 7% gains for Level 4 and Level 3 when the number of UEs is 15, demonstrating that Level 4 is more sensitive for the compression noise. In addition, Level 4 with WF-BD exhibits the slowest SE degradation, presenting the potential for scaling in dense networks.

Fig. 14 illustrates the sum SE with MMSE/L-MMSE combining under varying values of α. As α increases, both Level 3 and Level 4 experience a decline in SE due to increased

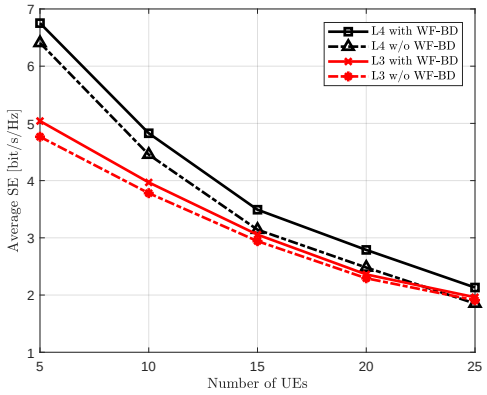


Fig. 13: Average SE under different number of UEs with $L = 12$ and $\tau_p = 5$.

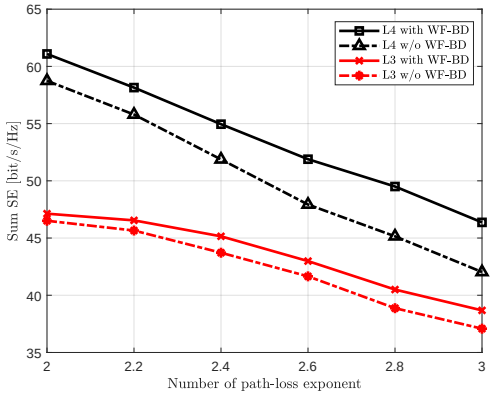


Fig. 14: Sum SE under different number of path-loss exponent with $L = 12$ and $J = 5$.

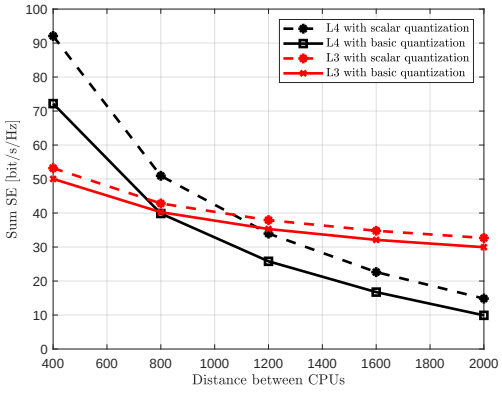


Fig. 15: Sum SE under different distance between CPUs with $J = 5$.

path-loss. Notably, the SE of Level 3 decreases less sharply compared to that of Level 4. Furthermore, the proposed WF-BD scheme effectively mitigates the SE degradation caused by large-scale fading, and its benefit becomes more pronounced as the path-loss exponent grows.

Fig. 15 shows the sum SE under different distance between CPUs with different quantization schemes. The first observation is that sclar quantization scheme can achieve about 8 bit/s/Hz and 2 bit/s/Hz SE gains for Level 4 and Level 3.

But the overall trend remains unchanged, i.e., Level 4 still declines significantly more than Level 3. This reflects that Level 4 is more sensitive to the quantization noise resulting from more severe compression. While scalar quantization can help alleviate the SE degradation to some extent, it cannot fully eliminate the addition noise especially for Level 4.

VII. Conclusion

In this paper, we conducted an investigation into the uplink performance of CF massive MIMO with mmWave fronthaul and backhaul. We considered four levels of multiple CPUs cooperation schemes inspired by [10], aiming to balance the degreenes significantly more than other levels of centralized coordination with the signal overhead on the backhaul link. Furthermore, we derived novel closed-form SE expressions for Level 2 and Level 3 using the MR combining. In addition, we also proposed a bisection search method to determine the valuable BCA factor for maximizing the uplink sum SE. Through simulation results, we examined the impact of mmWave fronthaul and backhaul, comparing the SE performance across various CPU cooperation levels. Notably, mmWave backhaul amplifies the compression noise caused by mmWave fronthaul, thereby having a more significant effect on the system. Despite this, the proposed multi-CPU cooperation schemes still outperform the non-cooperative scheme (Level 1), with the performance gains increasing as the number of CPUs grows. Moreover, as the number of APs or the distance between CPUs increase, Level 4 progressively loses its advantage due to the intensified compression noise, making Level 3 a more competitive option under these conditions. Based on these findings, we recommend allocating higher bandwidth for the backhaul link, as well as considering the usage of Level 3 when dealing with a large number of APs or significant distances between CPUs.

References

[1] E. Shi *et al.*, "RIS-aided cell-free massive MIMO systems for 6G: Fundamentals, system design, and applications," in *Proc. IEEE*, vol. 112, no. 4, pp. 331-364, Apr. 2024.

[2] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," in *Foundations and Trends in Signal Processing*, vol. 11, nos. 3-4. 2017, pp. 154-655.

[3] G. Femenias and F. Riera-Palou, "From cells to freedom: 6G's evolutionary shift with cell-free massive MIMO," *IEEE Trans. Mob. Comput.*, vol. 24, no. 2, pp. 812-829, Feb. 2025.

[4] J. Zhang, J. Zhang, D. W. K. Ng, S. Jin, and B. Ai, "Improving sum-rate of cell-free massive MIMO with expanded compute-and-forward," *IEEE Trans. Signal Process.*, vol. 70, no. 12, pp. 202-215, Dec. 2021.

[5] Z. Wang, J. Zhang, B. Ai, C. Yuen, and M. Debbah, "Uplink performance of cell-free massive MIMO with multi-antenna users over jointly correlated Rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7391-7406, Sep. 2022.

[6] J. Zhang *et al.*, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637-1660, Aug. 2020.

[7] Z. Li, J. Hu, X. Li, H. Zhang, and G. Min, "Dynamic AP clustering and power allocation for CF-MMIMO-enabled federated learning using multi-agent DRL," *IEEE Trans. Mob. Comput.*, early access, 2025.

[8] L. Bai, J. Xu, J. Wang, R. Han and J. Choi, "Efficient hybrid transmission for cell-free systems via NOMA and multiuser diversity," *IEEE Trans. Mob. Comput.*, vol. 24, no. 4, pp. 3359-3371, Apr. 2025.

[9] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1086-1100, Apr. 2021.

[10] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77-90, Jan. 2019.

[11] Z. Wang, J. Zhang, H. Q. Ngo, B. Ai, and M. Debbah "Uplink precoding design for cell-free massive MIMO with iteratively weighted MMSE," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1646 - 1664, Mar. 2023.

[12] Q. Sun *et al.*, "Uplink performance of hardware-impaired cell-free massive MIMO with multi-antenna users and superimposed pilots," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6711-6726, Nov. 2023.

[13] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1-6.

[14] H. Zhang, H. Li, T. Liu, L. Dong, and H, Wang, "On the total energy consumption of scalable cache-aided multi-CPU cell-free massive MIMO systems," *Digital Signal Processing*, vol. 154, pp. 104669, 2024.

[15] R. P. Antonioli, I. M. Braga, G. Fodor, Y. C. B. Silva, and W. C. Freitas, "Mixed coherent and non-coherent transmission for multi-CPU cell-free systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2023, pp. 1068-1073.

[16] T. Murakami, N. Aihara, A. Ikami, Y. Tsukamoto, and H. Shinbo, "Analysis of CPU placement of cell-free massive MIMO for user-centric RAN," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp.*, 2022, pp. 1-7.

[17] S. Kim, S. Ahn, J. Park, J. Youn, Y. Kwon, and S. Cho, "Revisiting the coverage boundary of multi-CPU cell-free massive MIMO: CPU cooperation aspect," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2023, pp. 1022-1028.

[18] S. Kim, S. Ahn, J. Park, J. Youn, Y. Kwon, and S. Cho, "CPU-cooperative power control scheme for scalable cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, early access, Jun. 2024.

[19] J. G. Andrews *et al.*, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065-1082, Jun. 2014.

[20] N. Rajapaksha, K. B. S. Manosha, N. Rajatheva, and M. Latva-aho, "Unsupervised learning-based joint power control and fronthaul capacity allocation in cell-free massive MIMO with hardware impairments," *IEEE Trans. Wireless Commun. Lett.*, vol. 12, no. 7, pp. 1159-1163, July 2023.

[21] I.-s. Kim, M. Bennis, and J. Choi, "Cell-free mmWave massive MIMO systems with low-capacity fronthaul links and low-resolution ADC/DACs," *IEEE Trans. Veh. Technol.*, vol. 71, no. 10, pp. 10512-10526, Oct. 2022.

[22] H. Masoumi and M. J. Emadi, "Cell-free massive MIMO system with limited fronthaul capacity and hardware impairments," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1038-1053, Feb. 2020.

[23] Z. Li, F. Göttsch, S. Li, M. Chen, and G. Caire, "Joint fronthaul load balancing and computation resource allocation in cell-free user-centric massive MIMO networks," *IEEE Trans. Wireless Commun.*, early access, Jun. 2024.

[24] M. Kim, I.-s. Kim, and J. Choi, "Meta-heuristic fronthaul bit allocation for cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11737-11752, Sept. 2024.

[25] F. J. Effenberger and D. Zhang, "WDM-PON for 5G wireless fronthaul," *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 94-99, Apr. 2022.

[26] M. Jiang *et al.*, "Wireless fronthaul for 5G and future radio access networks: Challenges and enabling technologies," *IEEE Wireless Commun.*, vol. 29, no. 2, pp. 108-114, Apr. 2022.

[27] U. Demirhan and A. Alkhateeb, "Enabling cell-free massive MIMO systems with wireless millimeter wave fronthaul," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9482-9496, Nov. 2022.

[28] S. Elhoushy, M. Ibrahim, and W. Hamouda, "Downlink performance of CF massive MIMO under wireless-based fronthaul network," *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 2632-2653, May 2023.

[29] C. Diaz-Vilor, A. Lozano, and H. Jafarkhani, "Cell-free UAV networks with wireless fronthaul: Analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 2054-2069, Mar. 2024.

[30] M. Ibrahim, S. Elhoushy, and W. Hamouda, "Uplink performance of mmWave-fronthaul cell-free massive MIMO systems," *IEEE Veh. Technol.*, vol. 71, no. 2, pp. 1536-1548, Feb. 2022.

[31] M. R. Akdeniz *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164-1179, Jun. 2014.

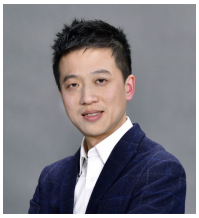
[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory.*, NJ, USA: Wiley, 2012.

[33] A. El Gamal and Y.-H. Kim, *Network Information Theory.*, U.K.: Cambridge Univ. Press, 2011.

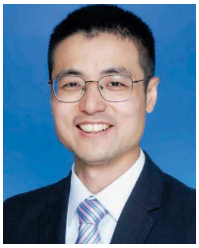
[34] M. Bashar, P. Xiao, R. Tafazolli, K. Cumanan, A. G. Burr, and E. Björnson, "Limited-fronthaul cell-free massive MIMO with local MMSE receiver under Rician fading and phase shifts" *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 1934-1938, Sept. 2021.

[35] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, vol. 159. New York, NY, USA: Springer, 2012.
[36] A. Lin, "Binary search algorithm," *WikiJ. Sci.*, vol. 2, no. 1, pp. 1-13, Jan. 2019.
[37] R. Wang, Y. Yang, B. Makki, and A. Shamim, "A wideband reconfigurable intelligent surface for 5G millimeter-wave applications," *IEEE Trans. Antennas Propag.*, vol. 72, no. 3, pp. 2399-2410, Mar. 2024.

Feiyang Li received the B.S. degree from the College of Information Science and Technology, Nantong University, Nantong, China, in 2021. He is currently pursuing the Ph.D. degree with Nantong University. His research interests include massive MIMO systems, backscatter communication, and performance analysis of wireless communication systems.



Qiang Sun (Member, IEEE) received the Ph.D. degree in communications and information systems from Southeast University, Nanjing, China, in 2014. He was a Visiting Scholar with the University of Delaware, Newark, DE, USA, in 2016. He is currently a Professor with the School of Information Science and Technology, Nantong, China. His research interests include deep learning and wireless communications. He was a member of Technical Program Committee and a reviewer for a number of IEEE conferences/journals.



Jiayi Zhang (Senior Member, IEEE) received the B.Sc. and Ph.D. degree of Communication Engineering from Beijing Jiaotong University, China in 2007 and 2014, respectively.

Since 2016, he has been a Professor with School of Electronic and Information Engineering, Beijing Jiaotong University, China. From 2014 to 2016, he was a Postdoctoral Research Associate with the Department of Electronic Engineering, Tsinghua University, China. From 2014 to 2015, he was also a Humboldt Research Fellow in Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Germany. From 2012 to 2013, he was a visiting scholar at the Wireless Group, University of Southampton, United Kingdom. His current research interests include cell-free massive MIMO, reconfigurable intelligent surface (RIS), communication theory and applied mathematics. Dr. Zhang received the Best Paper Awards at the WCSP 2017 and IEEE APCC 2017, the URSI Young Scientist Award in 2020, and the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2020. He was recognized as an exemplary reviewer of the IEEE COMMUNICATIONS LETTERS in 2015-2017. He was also recognized as an exemplary reviewer of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2017-2019. He was the Lead Guest Editor of the special issue on "Multiple Antenna Technologies for Beyond 5G" of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and an Editor for IEEE COMMUNICATIONS LETTERS from 2016-2021. He currently serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE ACCESS and IET COMMUNICATIONS.

Cunhua Pan (Senior Member, IEEE) is a full professor in Southeast Uni versity. His research interests mainly include recon f igurable intelligent surfaces (RIS), AI for Wireless, near field communications and sensing, and inte grated sensing and communications. He has pub lished over 200 IEEE journal papers. His papers got over 18600 Google Scholar citations with H-index of 69. He is Clarivate Highly Cited researcher. He is/was an Editor of IEEE Transaction on Communications, IEEE Transactions on Vehicular Technology, IEEE Wireless Communication Letters, and IEEE Communications Letters. He serves as the (leading) guest editors for IEEE Journal on Selected Areas in Communications, IEEE Journal of Selected Top ics in Signal Processing, IEEE Internet of Things, IEEE Vehicular Technology Magazine, IEEE Internet of Things Magazine, IEEE Transactions on Green Communications and Networking. He received the IEEE ComSoc Leonard G. Abraham Prize in 2022, IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2022, IEEE ComSoc Fred W. Ellersick Prize in 2024, IEEE ComSoc CTTC Early Achievement Award in 2024, IEEE ComSoc SPCC Early Achievement Award in 2024, and IEEE WCSP 2022 best paper award. He supervised one Phd Student to win the IEEE Signal Processing Society Best Phd Dissertation Award.

Kai-Kit Wong (Fellow, IEEE) (M'01-SM'08-F'16) received the BEng, the MPhil, and the PhD degrees, all in Electrical and Electronic Engineering, from the Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. After graduation, he took up academic and research positions at the University of Hong Kong, Lucent Technologies, Bell-Labs, Holmdel, the Smart Antennas Research Group of Stanford University, and the University of Hull, UK. He is Chair in Wireless Communications at the Department of Electronic and Electrical Engineering, University College London, UK. His current research centers around 6G and beyond mobile communications. He is Fellow of IEEE and IET. He served as the Editor-in-Chief for IEEE Wireless Communications Letters between 2020 and 2023.