Digital Communications and Networks 8 (2022) 695–719

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.keaipublishing.com/dcan

A survey on user-centric cell-free massive MIMO systems

Shuaifei Chen [a,b], Jiayi Zhang [a,b,*], Jing Zhang [a,b], Emil Björnson [c,d], Bo Ai [b,e]

[a] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, 100044, China
[b] Frontiers Science Center for Smart High-speed Railway System, Beijing Jiaotong University, Beijing, 100044, China
[c] Department of Computer Science, KTH Royal Institute of Technology, SE-16 440, Kista, Sweden
[d] Department of Electrical Engineering (ISY), Linköping University, SE-58 183, Linköping, Sweden
[e] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, 100044, China

ARTICLE INFO

Keywords:
6G network
User-centric cell-free network
Massive multiple-input multiple-output

ABSTRACT

The mobile data traffic has been exponentially growing during the last several decades. This was enabled by the densification of the network infrastructure in terms of increased cell density (i.e., Ultra-Dense Network (UDN)) and/or the increased number of active antennas per Access Point (AP) (i.e., massive Multiple-Input Multiple-Output (mMIMO)). However, neither UDN nor mMIMO will meet the increasing demand for the data rate of the Sixth Generation (6G) wireless communications due to the inter-cell interference and large quality-of-service variations. Cell-Free (CF) mMIMO, which combines the best aspects of UDN and mMIMO, is viewed as a key solution to this issue. In such systems, each User Equipment (UE) is served by a preferred set of surrounding APs cooperatively. In this paper, we provide a survey of the state-of-the-art literature on CF mMIMO. As a starting point, the significance and the basic properties of CF mMIMO are highlighted. We then present the canonical framework to discuss the essential details (i.e., transmission procedure and mathematical system model). Next, we provide a deep look at the resource allocation and signal processing problems related to CF mMIMO and survey the up-to-date schemes and algorithms. After that, we discuss the practical issues in implementing CF mMIMO and point out the potential future directions. Finally, we conclude this paper with a summary of the key lessons learned in this field.

1. Introduction

The performance of a mobile network is primarily quantified by the data rates that it can deliver to its users. Since there are a multitude of User Equipments (UEs) distributed over the coverage area, each experiencing unique propagation conditions, the rates that can be supported are highly user- and location-dependent. For example, the IMT-2020, requirements for the Fifth Generation (5G) technology specify the downlink peak data rates that are 200 times larger than the so-called user-experienced data rates that should be guaranteed to 95% of the users in the designated coverage area [1]. When reducing the network performance into a single metric (e.g., to enable network dimensioning), it is common to consider the area traffic capacity, which is measured as the total data rate of all active users divided by the coverage area [1]. When the network infrastructure is evolved to improve the area traffic capacity, the average data rates of the individual users will naturally increase, but it might have little impact on the user-experienced data rates, which are determined by the worst-case situations in the propagation environment. It is the user-experienced rates that determine which applications can be utilized without interruption in the system, not the average or peak rates. Hence, to enable the digitalization of society with a high perceived user fairness and consistent experience, the future network evolution should focus on improving the user-experienced data rates.

As noticed by Cooper [2], the area traffic capacity of cellular networks is determined by the available bandwidth, physical-layer technology, and cell density. In the past several decades, the vast majority of the improvements in traffic capacity is due to the densification of the network infrastructure in terms of increased cell density. This is much in line with the original cellular philosophy [3]: the coverage area is divided into cells served by different Access Points (APs), so that the number of active users per cell is manageable for the AP. Cellular networks were originally designed for voice services (i.e., mobile telephony), which are characterized by requiring a certain Signal-to-Noise Ratio (SNR) to give an

* Corresponding author.
E-mail addresses: shuaifeichen@bjtu.edu.cn (S. Chen), jiayizhang@bjtu.edu.cn (J. Zhang), 19120167@bjtu.edu.cn (J. Zhang), emilbjo@kth.se (E. Björnson), boai@bjtu.edu.cn (B. Ai).

https://doi.org/10.1016/j.dcan.2021.12.005
Received 29 May 2021; Received in revised form 16 December 2021; Accepted 23 December 2021
Available online 30 December 2021
2352-8648/© 2021 Chongqing University of Posts and Telecommunications. Publishing Services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

acceptable voice quality. If the SNR is below a threshold set by the codec, the voice is distorted and the call will eventually be dropped. As long as the SNR is above the threshold, the sound is distortion-free and thus the user experience is identical irrespective of how far above the threshold the SNR is. Hence, the first generations of cellular networks could be dimensioned based on two principles: first, to provide SNRs above the threshold almost everywhere in the coverage area; then, to densify in regions where the number of active UEs is above what the APs can handle in the peak hours.

The situation is much different since mobile broadband became the dominant service in cellular networks, because the data rate increases continuously with the SNR [4], up to the point where the maximum Spectral Efficiency (SE) is achieved. Hence, cell densification has two positive impacts on the area traffic capacity of mobile broadband services: more UEs can be simultaneously active in the network, and their SNRs increase, which leads to a higher rate per UE. Current networks consist of a mix of macro cells, micro cells, and small cells [5–8].

5G features an additional type of densification: a large number of active antennas per AP, which is known as massive Multiple-Input Multiple-Output (mMIMO) [9,10]. This technology makes the radiation pattern of the APs highly adaptable and more directive, so that a larger fraction of the transmitted power reaches the region around the receiver, while there is less interference at the undesired locations. Moreover, the technology allows for spatial multiplexing of UEs within each cell if the UEs are located in sufficiently different parts of the cell [11,12]. To enable efficient interference suppression in the spatial domain, a characteristic feature of mMIMO is that the AP has more antennas than the active UEs in the cell. Broadly speaking, mMIMO has the same two positive impacts on the area traffic capacity as cell densification, but they are achieved differently. The benefit of mMIMO compared to cell densification is that fewer APs are required to achieve a certain area capacity, while the drawback is that each AP is equipped with more complicated hardware. So far, 5G makes use of small cells in millimeter-wave bands and mMIMO in sub-6 GHz bands.

The densification is expected to continue beyond 5G [13], but both cell densification and mMIMO have fundamental limitations. As the cell area shrinks, the average SNR within a cell will improve, but the number of interfering cells also grows, which will eventually dominate. It is shown in Ref. [14] that this effect is noticeable already when there are 10 APs per km^2. This result relies on the assumption that all APs are transmitting simultaneously. In the so-called ultra-dense network regime [15,16], where there are many more APs than active UEs, only a random subset of the APs will have UEs to serve at any given point in time. The inactive APs will not cause interference in this regime, but since most APs will be idle most of the time, the required network infrastructure is utilized very inefficiently. When it comes to mMIMO, it is a highly scalable technology in terms of the ability to multiplex many users spatially [17], and one can increase the array dimensions proportionally to the number of UEs that need to be served. However, the technology is rather inefficient in overcoming the large SNR differences that the UEs experience within macro and micro cells. In summary, both cell densification and mMIMO might be well suited for increasing the peak and average rates in future cellular networks. Still, the user-experienced data rates will remain modest due to inter-cell interference and large SNR variations.

1.1. Cell-free mMIMO: the best of two worlds?

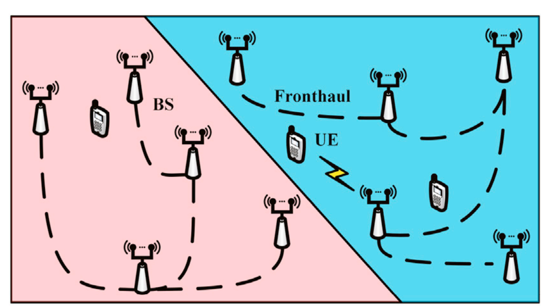
Cell-Free mMIMO (CF mMIMO) is a new technology that basically combines the best aspects of ultra-dense cellular networks with the cellular mMIMO technology to overcome their respective weaknesses [4]. The name was coined in Ref. [18] and refers to a network with many more APs than UEs and where the APs are cooperating to serve the UEs through coherent joint transmission and reception. One way to picture it is to take a network containing a single mMIMO array, dismantle the array, and deploy the individual antennas at different locations while keeping the same transmission/reception algorithms. When serving a given UE, the distributed antennas will then transmit each data signal with different power and phase-shifts, so they reach the intended UE synchronously, thereby reinforcing each other. Similarly, the received signals at different distributed antennas are co-processed to extract the data from each UE. Another way to view the creation of the technology is to start from an ultra-dense cellular network, connect the APs to form a virtual distributed mMIMO array, as illustrated in Fig. 1(a), and then utilize (roughly) the same transmission algorithms as a conventional mMIMO array would do.

Irrespective of the direction from which one approaches the CF mMIMO technology, the main properties are that there are many geographically distributed APs, but the coverage area is not divided into disjoint cells. Each UE is served by all the surrounding APs, as illustrated in Fig. 1(b). This mMIMO processing resolves the interference situation that limits conventional ultra-dense networks and leads to a network free from cells. Moreover, by having many distributed AP antennas instead of few APs with large antenna arrays, the large SNR variations that limit the efficiency of the conventional cellular mMIMO are effectively mitigated. The original motivation behind CF mMIMO was to design a new network infrastructure capable of providing uniform data rates in the coverage area [18], that is, to concentrate on improving the user-experienced data rates, instead of the average or peak rates which are already quite high in contemporary networks.

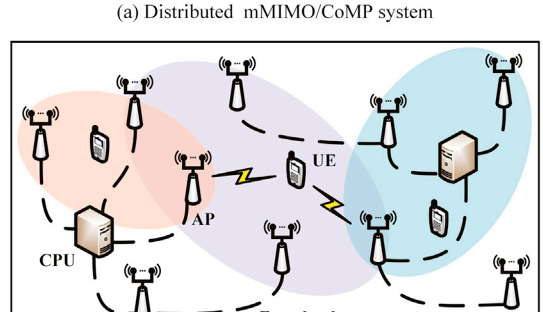
Since each UE will only be influenced by the signals from the closest surrounding APs, a CF mMIMO system can also be viewed as a user-centric network [4,19–21]. As illustrated in Fig. 1(b), each UE is served by a unique set of surrounding APs. To facilitate the cooperation between the neighboring APs flexibly, the technology is conceived to make use of a Cloud Radio Access Network (C-RAN) infrastructure [22]. More precisely, the APs are connected via the so-called fronthaul connections to one or multiple edge-cloud processors, which are called Central Processing Units (CPUs) in the CF mMIMO literature [18]. The backhaul connections can either be fully wired (e.g., using optical fiber cables) or partially wireless (e.g., using fixed microwave links).

1.2. Related technologies

The vision of serving UEs using multiple distributed APs has been around for a few decades. For example, Wyner described in Ref. [23] how



(a) Distributed mMIMO/CoMP system



(b) CF mMIMO system

Fig. 1. Comparison of distributed mMIMO/CoMP and CF mMIMO.

one can untangle interfering uplink signals by joint detection at the neighboring APs. One can view this as a system of linear equations where each unknown variable is the information transmitted by one UE, and each equation is the signal received at one AP. A single-antenna AP can only identify one UE signal, but if the neighboring APs cooperate, they can jointly identify as many UE signals as APs. A downlink counterpart of this concept was introduced by Shamai and Zaidel in Ref. [24] from 2001. Early embodiments of this technology have been called *Distributed Wireless Communication System* [25] and *Network MIMO* [26]. Other prominent papers in this period are [27–37] and the survey article [38]. 3GPP called these technologies *Coordinated MultiPoint* (CoMP) [39,40]. The general premise was to evolve an existing cellular network by adding the cooperation between the neighboring APs to reduce the inter-cell interference, not to build a cell-free network from scratch as is the vision with CF mMIMO. Two main distinguishing factors between these early works and CF mMIMO are the operating regime with more APs than UEs and the physical-layer operation inspired by the recent advancements in the mMIMO field. For example, perfect Channel State Information (CSI) was generally assumed in the Network MIMO literature, and the methodology for analyzing the achievable data rates under imperfect CSI was largely missing at that time. When CoMP algorithms were analyzed under practical conditions, the gains were surprisingly low [40]. In practice, only the closest APs can acquire reliable CSI. Thus the system operation must be made robust against CSI imperfections, and there must be a resource-efficient way to acquire CSI. The mMIMO methodology provides these missing pieces.

The intended use case of the CoMP technology was to take an existing cellular network and divide the APs into disjoint clusters [41–44], which effectively creates a cellular network with distributed antennas within each cell cluster. The significant difference between distributed mMIMO/CoMP and CF mMIMO is illustrated in Fig. 1 and Table 1. The CoMP technology reduces the SNR variations within each cluster but keeps the cellular structure. Thus UEs at the edges of a cell cluster are affected by interference from neighboring cell clusters. This is not the case in a CF mMIMO network, where every UE is served by all the surrounding APs (see Table 2).

1.3. Contributions of this survey

CF mMIMO evolves from multiple existing techniques, like the Ultra-Dense Network (UDN), CoMP, Network MIMO, and mMIMO. The convergence of these different technologies brings about the novel strategies and procedures to meet the stringent requirements of the Sixth Generation (6G) network in terms of the high user-experienced data rates and ubiquitous coverage which necessitates a survey of the existing work on this topic and addressing the future research directions.

Table 1
Comparison of cellular mMIMO, distributed mMIMO/CoMP, and CF mMIMO.

Technology	Cellular mMIMO	Distributed mMIMO/CoMP	CF mMIMO
Coverage	Small	Medium	Large
Clustering	Network-centric	Network-centric	User-centric
	Disjoint	Disjoint	Partially overlapping
	Fixed	Fixed	Dynamic
A UE served by	One BS	A few APs	All surrounding APs
CSI for decoding	Instantaneous	Instantaneous	Instantaneous or statistical
Fronthaul load	–	Large	Small
Synchronization	Uncritical	Critical	Critical
User-experience rate	Low	Medium	Large

Table 2
Important abbreviations.

Abbreviation	Definition	Abbreviation	Definition
5G	fifth generation	IoT	Internet of Things
6G	Sixth generation	ISAC	Integrated sensing and communication
ADC	Analog-to-digital converter	LMMSE	Linear minimum mean-squared error
ANN	Artificial neural network	L-MMSE	Local minimum mean-squared error
AoA	Angle-of-arrival	LoS	Line-of-sight
AoD	Angles-of-departure	LP-MMSE	Local partial minimum mean-squared error
AP	Access point	L-RZF	Local regularized zero-forcing
APO	AP switch On/Off	LS	Least-square
C&F	Compute-and-Forward	LSFD	Large-scale fading decoding
CAP	Compress-after-precoding	MEC	Mobile edge computing
CBDNet	Convolutional blind denoising network	ML	Machine learning
CDF	Cumulative distribution function	mMIMO	Massive multiple-input multiple-output
CF	Cell-free	MMSE	Minimum mean-squared error
CFE	Compress-forward-estimate	MR	Maximum ratio
C-MMSE	Centralized minimum mean-squared error	MSE	Mean-squared error
CoMP	Coordinated multipoint	NLoS	Non line-of-sight
CPU	Central processing unit	NMSE	Normalized mean-squared error
CP	Cyclic prefix	OFDM	Orthogonal frequency division multiplexing
C-RAN	Cloud radio access network	OTA	Over-the-air
CS	Compressive sensing	PAC	Precoding-after-compress
CSI	Channel state information	PA-MMSE	Phase-aware minimum mean-squared error
DCNN	Deep convolutional neural network	P-FZF	Partial full-pilot zero-forcing
DFRC	Dual-functional radar-communication	P-MMSE	Partial minimum mean-squared error
DFT	Discrete Fourier transform	PWP-FZF	Protective weak partial full-pilot zero-forcing
E-C&F	Expanded compute-and-forward	QoS	Quality-of-service
ECF	Estimate-compress-forward	RAU	Radio access unit
EE	Energy efficiency	RZF	Regularized zero-forcing
EMCF	Estimate-multiply-compress-forward	SCA	Successive convex approximation
EMCFW	Estimate-multiply-compress-forward-weight	SE	Spectral efficiency
EW-MMSE	Element-wise minimum mean-squared error	SIC	Successive interference cancelation
FDD	Frequency-division duplex	SINR	Signal-to-interference-and-noise ratio
FFDNet	Flexible denoising convolutional neural network	SNR	Signal-to-noise ratio
FL	Federated learning	SOCP	Second-order cone program
FZF	Full-pilot zero-forcing	TDD	Time-division duplex
GP	Geometric programming	TOA	Time-of-arrival
GPS	Global position system	UatF	Use-and-then-forget
HI	Hardware impairments	UDN	Ultra-dense network
ICA	Independent component analysis	UE	User equipment
IoE	Internet of Everything	ZF	Zero-forcing

In this survey, we first describe in detail the technical foundations of CF mMIMO by giving a brief tutorial in a canonical framework. This serves for a better understanding of the extensive survey on the state-of-the-art schemes and algorithms for resource allocation, signal processing, practical implementation, and future research directions on this topic. To the best of the authors' knowledge, there has been no comprehensive survey on CF mMIMO available in the literature. Although there are three survey and tutorial articles [4,45,46] available, they all missed some parts of the holistic overview. Specifically, [4,45] provided tutorials on CF mMIMO while they lack a comprehensive survey on the applied schemes and algorithms in [45], and gave an early survey while extensive research has been conducted after that. This motivates this survey to review the up-to-date work on CF mMIMO to provide a starting point for anyone who wants to conduct research on this topic.

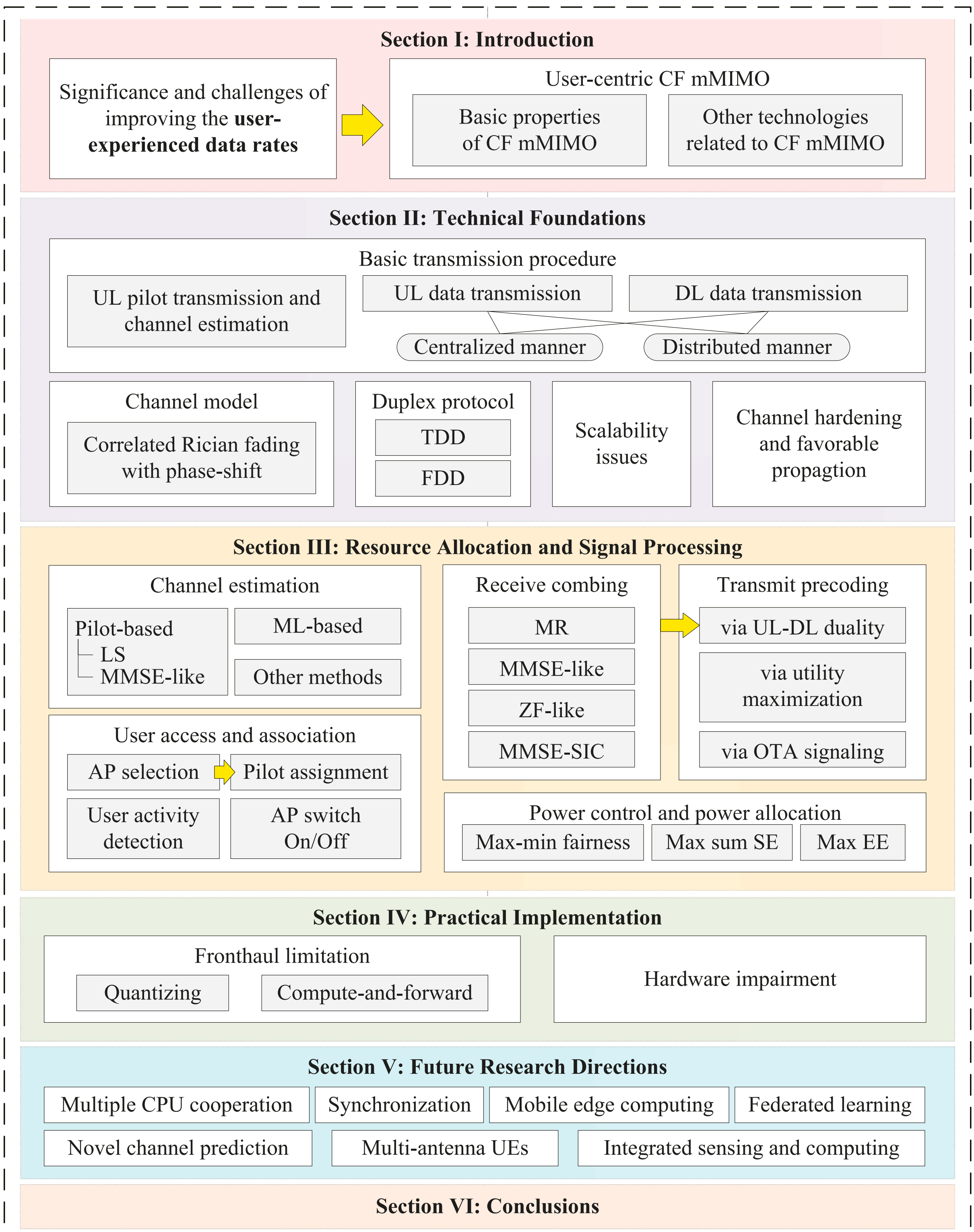


Fig. 2. Survey roadmap.

1.4. Paper outline

The remainder of this paper is organized as follows. Section 2 introduces the technical foundations of CF mMIMO, where the transmission procedure and mathematical system model are discussed. Section 3 gives a comprehensive survey on resource allocation and signal processing while the practical issues in implementing CF mMIMO are discussed in Section 4. Then, potential future directions of CF mMIMO research are highlighted in Section 5. Finally, this paper is concluded in Section 6 with a summary of the key lessons learned in this field. The overall roadmap of this paper is illustrated in Fig. 2.

1.5. Notation

Boldface lowercase letters, $\mathbf{x}$, denote column vectors and boldface uppercase letters, $\mathbf{X}$, denote matrices. The superscripts $^{\mathrm{T}}$, *, and $^{\mathrm{H}}$ denote transpose, conjugate, and conjugate transpose, respectively. The $n \times n$ identity matrix is $\mathbf{I}_n$. We use $\triangleq$ for definitions and $\mathrm{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ for a block-diagonal matrix with the square matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$ on the diagonal. The multi-variate circularly symmetric complex Gaussian distribution with correlation matrix $\mathbf{R}$ is denoted $\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R})$. The expected value of $\mathbf{x}$ is denoted as $\mathbb{E}\{\mathbf{x}\}$. We use $|\mathcal{A}|$ to denote the cardinality of the set $\mathcal{A}$.

2. Technical foundations

Having explained the basic motivation and properties of CF mMIMO in the previous section, we will now take a look at its fundamental technical components and details. We first present the basic transmission procedure of CF mMIMO systems. We then discuss the mathematical system model, including the models of the fading channels, duplex protocols, scalability issues, and channel hardening and favorable propagation.

2.1. Basic transmission procedure

We consider a representative CF mMIMO system consisting of K single-antenna UEs and L APs, each equipped with N antennas. As illustrated in Fig. 1(b), all APs are connected to a CPU in an arbitrary fashion with the fronthaul connections. These connections facilitate the cooperation between the APs, such as the coherent joint transmission of data signals to the UEs and the coherent joint reception of data signals from the UEs. The system could operate either in Time-Division Duplex (TDD) mode or in Frequency-Division Duplex (FDD) mode, which is further discussed in Section 2.3. For now, all APs and UEs are assumed to be operated in TDD mode. The propagation channels vary over time and frequency, which we describe using a block fading model [4]. More precisely, the time-frequency grid is divided into coherence blocks of τ_c channel uses for which the channel is constant and frequency flat. Each coherence block is divided into three phases: τ_p channel uses for uplink channel estimation, τ_u and τ_d for uplink and downlink data, respectively, such that $\tau_c = \tau_p + \tau_u + \tau_d$. Note that this block-fading model is an abstraction of the practical multi-carrier modulation model used in Orthogonal Frequency Division Multiplexing (OFDM), and we consider it to make sure that the main concepts of CF mMIMO are not hidden under the more complicated notation that OFDM requires. However, one can still map a practical OFDM system to this block-fading system by regarding several subcarriers which compose the available bandwidth as an aforementioned coherence block. See [47, Sec. 2] for a concrete example. We denote by $\mathbf{h}_{kl} \in \mathbb{C}^N$ the channel response between AP l and UE k, which is a random realization in each coherence block of some stationary ergodic fading distribution. Different fading distributions will be discussed in Section 2.2.

2.1.1. Uplink pilot transmission and channel estimation

When the UEs have gained access to the network, they are assigned pilots which are used for channel estimation. We assume there are τ_p mutually orthogonal τ_p-length pilot signals, where τ_p is a constant independent of K. The pilot resources are limited due to the natural channel variations in the time and frequency domains. Thus we have $\tau_p < K$ in most practical scenarios, and the pilots must thus be reused between UEs. Different algorithms for the pilot assignment are surveyed in Section 3.4. For now, we denote by $t_k \in \{1, \ldots, \tau_p\}$ the index of the pilot assigned to UE k and call $\mathcal{S}_k = \{i : t_k = t_i\} \subset \{1, \ldots, K\}$ the subset of UEs sharing pilot t_k, including UE k. When the UEs in $\mathcal{S}_k$ transmit pilot t_k, the received signal

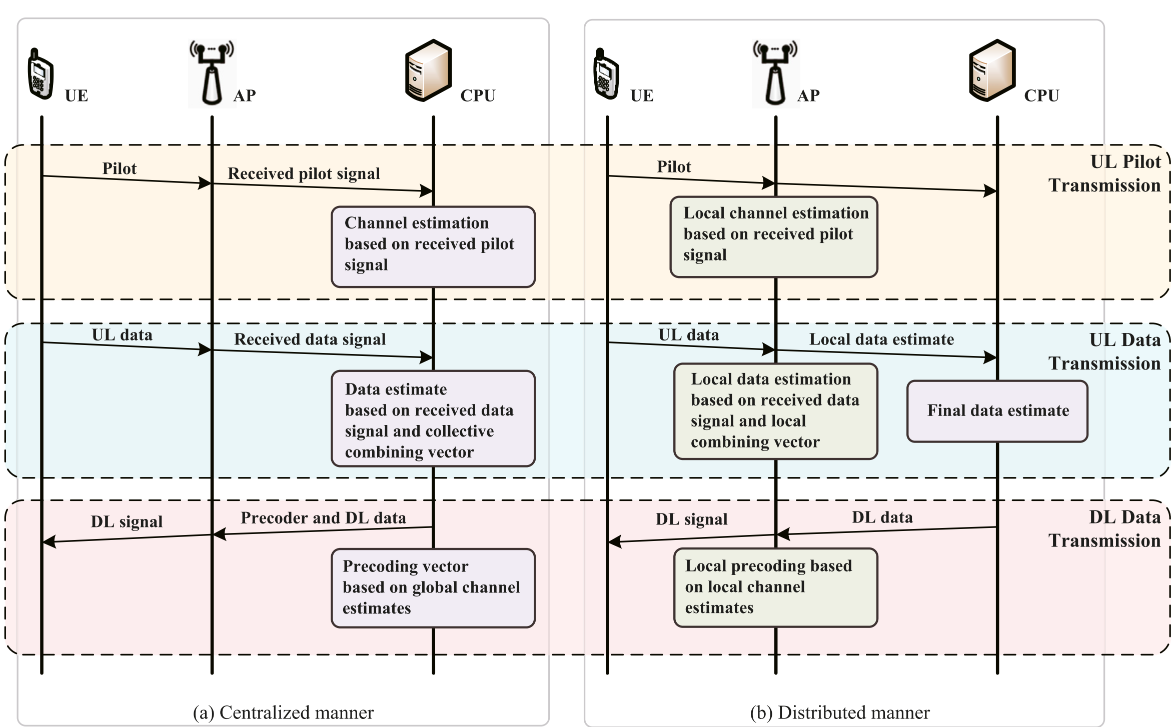


Fig. 3. Flow chart of three-stage transmission procedure for CF mMIMO.

$\mathbf{y}_{t_k l}^{\text{pilot}} \in \mathbb{C}^N$ after despreading at AP l is [4, Sec. 3]

$$\mathbf{y}_{t_k l}^{\text{pilot}} = \sum_{i \in \mathcal{S}_k} \sqrt{\tau_p p_i} \mathbf{h}_{il} + \mathbf{n}_{t_k l} \tag{1}$$

where p_i is the transmit power of UE i and $\mathbf{n}_{t_k l} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma_{\text{ul}}^2 \mathbf{I}_N)$ is the thermal noise.

By exploiting different levels of prior information regarding the fading distributions, different kinds of channel estimators could be utilized at the APs (or CPU) to estimate the channels based on (1), which are elaborated in Section 3.1. For now, we denote by $\hat{\mathbf{h}}_{kl} \in \mathbb{C}^N$ the estimate of channel $\mathbf{h}_{kl}$. As shown in Fig. 3, an AP can perform the channel estimation locally or delegate this task to the CPU by just handing over its received pilot signals. Since the pilots are reused between UEs, there will be interference between the pilot-sharing UEs. This will reduce the channel estimation quality, which makes the coherent transmission less effective. It also makes it harder to reject interference between the UEs sharing the same pilot since the AP cannot correctly separate their channels. This is the so-called *pilot contamination* phenomenon, which is particularly famous in the cellular mMIMO literature [11,48], but exists in any wireless system with non-orthogonal pilot transmissions.

2.1.2. Uplink data transmission

During uplink data transmission, the received signal $\mathbf{y}_l^{\text{ul}} \in \mathbb{C}^N$ at AP l is

$$\mathbf{y}_l^{\text{ul}} = \sum_{i=1}^{K} \mathbf{h}_{il} s_i + \mathbf{n}_l \tag{2}$$

where $s_i \in \mathbb{C}$ is the signal transmitted from UE i which is assumed to be transmitted with power p_i, and $\mathbf{n}_l \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma_{\text{ul}}^2 \mathbf{I}_N)$ is the noise. One can estimate s_k by properly combining the received signal $\mathbf{y}_l^{\text{ul}}$. For now, we let $\hat{s}_k$ denote the estimate of s_k and call $\mathbf{v}_{kl} \in \mathbb{C}^N$ the combining vector that AP l assigns to UE k. Due to the three-stage architecture of CF mMIMO, there can be two levels of cooperation among APs for combing design, i.e., centralized combining and distributed combining as shown in Fig. 3.

Centralized combining. In the first level, all APs forward their received data signals $\{\mathbf{y}_l^{\text{ul}} : l = 1, \ldots, L\}$ to the CPU, which performs channel estimation and data detection in a fully centralized fashion. Since no local signal processing is performed at APs, the CPU sees a collection of the received data signals

$$\mathbf{y}^{\text{ul}} = \sum_{i=1}^{K} \mathbf{h}_i s_i + \mathbf{n} \tag{3}$$

where $\mathbf{y}^{\text{ul}} = [(\mathbf{y}_1^{\text{ul}})^{\text{T}}, \ldots, (\mathbf{y}_L^{\text{ul}})^{\text{T}}]^{\text{T}} \in \mathbb{C}^{LN}$, $\mathbf{h}_i = [\mathbf{h}_{i1}^{\text{T}}, \ldots, \mathbf{h}_{iL}^{\text{T}}]^{\text{T}} \in \mathbb{C}^{LN}$, and $\mathbf{n} = [\mathbf{n}_1^{\text{T}}, \ldots, \mathbf{n}_L^{\text{T}}]^{\text{T}} \in \mathbb{C}^{LN}$. Based on all the collective channel estimates $\{\hat{\mathbf{h}}_k = [\hat{\mathbf{h}}_{k1}^{\text{T}}, \ldots, \hat{\mathbf{h}}_{kL}^{\text{T}}]^{\text{T}} \in \mathbb{C}^{LN} : k = 1, \ldots, K\}$, the CPU can select an arbitrary combining vector $\mathbf{v}_k \in \mathbb{C}^{LN}$ for UE k. Consequently, the data transmitted by UEs can be estimated.

We cannot compute the exact ergodic capacity of this setup due to the imperfect channel knowledge. However, we can rigorously analyze the performance by using a standard capacity lower bound [12] referred to as an achievable spectral SE. The following SE is achievable when using Minimum Mean-Squared Error (MMSE) channel estimation, which is introduced in Section 3.1.

Lemma 1. *When the MMSE channel estimates are available, an achievable uplink SE of UE k is*

$$\mathsf{SE}_k^{(\text{ul},1)} = \frac{\tau_u}{\tau_c} \mathbb{E}\left\{ \log_2\left(1 + \mathsf{SINR}_k^{(\text{ul},1)}\right) \right\} \tag{4}$$

where the instantaneous effective Signal-to-Interference-and-Noise Ratio (SINR) is

$$SINR_k^{(\text{ul},1)} = p_k \frac{\left|\mathbf{v}_k^{\text{H}} \hat{\mathbf{h}}_k\right|^2}{\sum\limits_{i=1, i \neq k}^{K} p_i \left|\mathbf{v}_k^{\text{H}} \hat{\mathbf{h}}_i\right|^2 + \mathbf{v}_k^{\text{H}} \left(\sum\limits_{i=1}^{K} p_i \mathbf{C}_i + \sigma_{\text{ul}}^2 \mathbf{I}_{LN} \right) \mathbf{v}_k} \tag{5}$$

with $\mathbf{C}_i = \text{diag}(\mathbf{C}_{i1}, \ldots, \mathbf{C}_{iL})$ *where* $\mathbf{C}_{il} = \mathbb{E}\{(\mathbf{h}_{il} - \hat{\mathbf{h}}_{il})(\mathbf{h}_{il} - \hat{\mathbf{h}}_{il})^{\text{H}}\}$ *is the error correlation matrix of* $\hat{\mathbf{h}}_{il}$.

Proof 1. *It follows from the proof of [12, Theo. 4.1].*

Distributed combining. In the second level, each AP can preprocess its signal by computing local estimates of the data and then passing them to the CPU for final decoding. With the local combining vector $\mathbf{v}_{kl}$, AP l computes its local estimate of s_k as

$$s_{kl} \triangleq \mathbf{v}_{kl}^{\text{H}} \mathbf{y}_l^{\text{ul}} = \mathbf{v}_{kl}^{\text{H}} \mathbf{h}_{kl} s_k + \sum_{i=1, i \neq k}^{K} \mathbf{v}_{kl}^{\text{H}} \mathbf{h}_{il} s_i + \mathbf{v}_{kl}^{\text{H}} \mathbf{n}_l \tag{6}$$

Any combining vector can be adopted in the above expression (6). Unlike the centralized combining, however, AP l can only use its own local channel estimates to design $\mathbf{v}_{kl}$.

The local estimates $\{s_{kl} : l = 1, \ldots, L\}$ are then sent to the CPU where they are linearly combined using the weights $\{a_{kl}: l = 1, \ldots, L\}$ to obtain $\hat{s}_k = \sum_{l=1}^{L} a_{kl}^* s_{kl}$ which is eventually used to decode s_k, as

$$\begin{aligned} \hat{s}_k &= \sum_{l=1}^{L} a_{kl}^* \mathbf{v}_{kl}^{\text{H}} \mathbf{h}_{kl} s_k + \sum_{l=1}^{L} a_{kl}^* \sum_{i=1, i \neq k}^{K} \mathbf{v}_{kl}^{\text{H}} \mathbf{h}_{il} s_i + \mathbf{n}_k' \\ &= \mathbf{a}_k^{\text{H}} \mathbf{f}_{kk} s_k + \sum_{i=1, i \neq k}^{K} \mathbf{a}_k^{\text{H}} \mathbf{f}_{ki} s_i + \mathbf{n}_k' \end{aligned} \tag{7}$$

where $\mathbf{f}_{ki} = [\mathbf{v}_{k1}^{\text{H}} \mathbf{h}_{i1} \ldots \mathbf{v}_{iL}^{\text{H}} \mathbf{h}_{kL}]^{\text{T}} \in \mathbb{C}^L$ is the receive-combined channel vector between UE k and each of the APs, $\mathbf{a}_k = [a_{k1}, \ldots, a_{kL}]^{\text{T}} \in \mathbb{C}^L$ is the weighting vector, $\mathbf{n}_k' = \sum_{l=1}^{L} a_{kl}^* \mathbf{v}_{kl}^{\text{H}} \mathbf{n}_l$ is the effective noise, and $\{\mathbf{a}_k^{\text{H}} \mathbf{f}_{ki} : i = 1, \ldots, K\}$ is the effective channel.

Since the CPU does not have the knowledge of the effective channel $\mathbf{a}_k^{\text{H}} \mathbf{f}_{kk}$, we utilize the well-considered *Use-and-then-Forget* (UatF) capacity bound [12, Theo. 4.4], where we use the channel estimates for combining and then effectively "forget" them before the signal detection to obtain the achievable SE.

Lemma 2. *An achievable uplink SE of UE k is*

$$\mathsf{SE}_k^{(\text{ul},2)} = \frac{\tau_u}{\tau_c} \log_2\left(1 + \mathsf{SINR}_k^{(\text{ul},2)}\right) \tag{8}$$

with the effective SINR given by

$$SINR_k^{(\text{ul},2)} = \frac{p_k \left|\mathbf{a}_k^{\text{H}} \mathbb{E}\{\mathbf{f}_{kk}\}\right|^2}{\sum\limits_{i=1}^{K} p_i \mathbb{E}\left\{\left|\mathbf{a}_k^{\text{H}} \mathbf{f}_{ki}\right|^2\right\} + p_k \left|\mathbf{a}_k^{\text{H}} \mathbb{E}\{\mathbf{f}_{kk}\}\right|^2 + \sigma_{\text{ul}}^2 \mathbf{a}_k^{\text{H}} \mathbf{\Lambda}_k \mathbf{a}_k} \tag{9}$$

where $\mathbf{\Lambda}_k = \text{diag}(\mathbb{E}\{\|\mathbf{v}_{k1}\|^2\}, \ldots, \mathbb{E}\{\|\mathbf{v}_{kL}\|^2\}) \in \mathbb{C}^{L \times L}$ *and the expectations are with respect to the channel estimates.*

Proof 2. *The proof is given in [49, Appe. A].*

The achievable SE above holds for any combining scheme. Unlike the achievable SE in Lemma 1, it holds for any channel estimator (not only for the MMSE estimator). The drawback with this bound is that it is only tight when $\mathbf{a}_k^{\text{H}} \mathbf{f}_{kk}$ is close to its mean value $\mathbf{a}_k^{\text{H}} \mathbb{E}\{\mathbf{f}_{kk}\}$, but this seems to be the case in many mMIMO setups [4].

The structure of (9) allows computing the deterministic weighting vector $\mathbf{a}_k$ that maximizes $\mathsf{SINR}_k^{(\text{ul},2)}$ as follows.

Corollary 1. *The effective SINR in (9) for UE k is maximized by*

$$\mathbf{a}_k = \left(\sum_{i=1}^{K} p_i \mathbb{E}\{\mathbf{f}_{ki} \mathbf{f}_{ki}^{\text{H}}\} + \sigma_{\text{ul}}^2 \mathbf{\Lambda}_k \right)^{-1} \mathbb{E}\{\mathbf{f}_{kk}\} \tag{10}$$

which leads to the maximum value

$$\mathsf{SINR}_k^{(\mathrm{ul},2)} = p_k\mathbb{E}\{\mathbf{f}_{kk}\}\left(\sum_{i=1}^{K} p_i\mathbb{E}\{\mathbf{f}_{ki}\mathbf{f}_{ki}^{\mathrm{H}}\} + \sigma_{\mathrm{ul}}^2\mathbf{\Lambda}_k - p_k\mathbb{E}\{\mathbf{f}_{kk}\}\mathbb{E}\{\mathbf{f}_{kk}^{\mathrm{H}}\}\right)^{-1}\mathbb{E}\{\mathbf{f}_{kk}\} \tag{11}$$

Proof 3. *It follows from [12, Lemm. B.10].*

The aforementioned approach is the so-called *Large-Scale Fading Decoding (LSFD)* [50]. Although this approach offers the highest SE among schemes with local combining at each AP, it requires sharing of statistical information sharing among the APs to design the optimized LSFD weights. Alternatively, the weight a_{kl} can be locally designed at each AP, such as $a_{kl} = \beta_{kl}^{\nu}$ with different exponents ν, where $\beta_{kl} \triangleq \mathrm{tr}(\mathbf{R}_{kl})/N$ is the large-scale fading coefficient that describes pathloss and shadowing and $\mathbf{R}_{kl}$ is the spatial correlation matrix that describes the spatial property of the channel. When $\nu = 0$, the CPU creates its estimate of the signal s_k from UE k by simply taking the average of the local estimates, as proposed in the early papers on CF mMIMO [51,52].

Note that the value of the uplink SE of UE k in CF mMIMO systems depends on the UE's combining vector $\mathbf{v}_k$. The choices of the combining vector in the CF mMIMO literature can be found in Section 3.3.

2.1.3. Downlink data transmission

Let $\mathbf{w}_{il} \in \mathbb{C}^N$ denote the precoder that AP l assigns to UE i. During downlink data transmission, the received signal at UE k is

$$y_k^{\mathrm{dl}} = \sum_{l=1}^{L} \mathbf{h}_{kl}^{\mathrm{H}} \sum_{i=1}^{K} \mathbf{w}_{il}\varsigma_i + n_k = \mathbf{h}_k^{\mathrm{H}} \sum_{i=1}^{K} \mathbf{w}_i\varsigma_i + n_k \tag{12}$$

where $\varsigma_i \in \mathbb{C}$ is the independent unit-power data signal intended for UE i (i.e., $\mathbb{E}\left\{\left|\varsigma_i\right|^2\right\} = 1$), $\mathbf{w}_k = [\mathbf{w}_{k1}^{\mathrm{T}}, \ldots, \mathbf{w}_{kL}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{LN}$ is the collective precoding vector, and $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\mathrm{dl}}^2)$ is the receiver noise. Normally, the collective precoding vector $\mathbf{w}_i$ is presented as

$$\mathbf{w}_i = \sqrt{\rho_i}\overline{\mathbf{w}}_i \tag{13}$$

where $\overline{\mathbf{w}}_i$ determines the spatial directivity of the transmission and satisfies $\mathbb{E}\{\|\overline{\mathbf{w}}_i\|^2\} = 1$ such that $\rho_i \geq 0$ is the transmit power allocated to UE i.

For a specific choice of precoding, the *hardening bound* is used to compute the downlink SE.

Lemma 3. *An achievable downlink SE of UE k is*

$$\mathsf{SE}_k^{(\mathrm{dl})} = \frac{\tau_d}{\tau_c}\log_2\left(1 + \mathsf{SINR}_k^{(\mathrm{dl})}\right) \tag{14}$$

with the effective SINR given by

$$SINR_k^{(\mathrm{dl})} = \frac{\left|\rho_k\mathbb{E}\{\mathbf{h}_k^{\mathrm{H}}\overline{\mathbf{w}}_k\}\right|^2}{\sum_{i=1}^{K}\rho_i\mathbb{E}\{|\mathbf{h}_k^{\mathrm{H}}\overline{\mathbf{w}}_i|^2\} - \rho_k\left|\mathbf{h}_k^{\mathrm{H}}\overline{\mathbf{w}}_k\right|^2 + \sigma_{\mathrm{dl}}^2} \tag{15}$$

and the expectation is with respect to the channel realizations.

Proof 4. *The proof is given in [12, Appe. C.3.6].*

In contrast to the uplink SEs of UE k that only depend on the UE's combining vector, the downlink SE depends on the precoding vectors of all UEs, i.e., $\{\mathbf{w}_i: i = 1, \ldots, K\}$. Consequently, the precoding vectors should be optimized jointly for all UEs instead of on a per-UE basis. Alternatively, one can utilize the following uplink-downlink duality result to obtain a good heuristic solution.

Lemma 4. *Let $\{\mathbf{v}_i: i = 1, \ldots, K\}$ and $\{p_i: i = 1, \ldots, K\}$ denote the set of combining vectors and transmit powers used in the uplink. If the normalized precoding vectors are selected as*

$$\overline{\mathbf{w}}_i = \frac{\mathbf{v}_i}{\sqrt{\mathbb{E}\{\|\mathbf{v}_i\|^2\}}} \tag{16}$$

then there exists a downlink power control policy ρ_i: $\forall i$ with $\sum_{i=1}^{K}\rho_i/\sigma_{\mathrm{dl}}^2 = \sum_{i=1}^{K}p_i/\sigma_{\mathrm{ul}}^2$ for which

$$\mathsf{SINR}_k^{(\mathrm{dl})} = \mathsf{SINR}_k^{(\mathrm{ul},2)}, \forall k \tag{17}$$

where $\mathsf{SINR}_k^{(\mathrm{dl})}$ is the effective SINR of UE k in the downlink and $\mathsf{SINR}_k^{(\mathrm{ul})}$ is the effective SINR of UE k in the uplink with distributed cooperation manner.

Proof 5. *The proof is given in [21, Appe.].*

The above lemma implies that the downlink precoders in CF mMIMO networks can be selected based on the uplink combiners as in (16). Consequently, an achievable downlink SE for UE k can be achieved by properly selecting the power control coefficients $\{\rho_i: \forall i\}$ and the normalized precoding vectors $\{\overline{\mathbf{w}}_i : \forall i\}$. Similar to that in the uplink, we consider two levels of cooperation among APs for the precoding design as shown in Fig. 3. At both levels, we assume that the APs delegate the task of downlink data encoding to the CPU.

Centralized precoding In the first level, the CPU uses the uplink channel estimates to compute the normalized precoding vectors $\{\overline{\mathbf{w}}_{il}\}$ by exploiting the channel reciprocity. Motivated by the uplink-downlink duality, we select the downlink precoding vectors according to (16). Once the precoding vectors are computed, they are used by the CPU to form the downlink signal of any given AP l, as

$$\mathbf{x}_l^{\mathrm{dl}} = \sum_{l=1}^{K}\sqrt{\rho_i}\overline{\mathbf{w}}_{il}\varsigma_i \tag{18}$$

which is sent to the AP via the fronthaul link for transmission.

Distributed precoding In the second level, AP l can locally select the precoding vector $\mathbf{w}_{il}$ on the basis of its local channel estimates $\{\hat{\mathbf{h}}_{il}\}$ instead of delegating the task to the CPU. In this case, only the downlink data signals $\{\varsigma_i\}$ are sent from the CPU to AP l in each coherence block.

The choices of the precoding vector in the CF mMIMO literature can be found in Section 3.3.

2.2. Channel model

Line-of-Sight (LoS) channels are widely considered to characterize the propagation channels, which generally contain many propagation paths. One is the direct path, and the others are paths where the signals are scattered on different objects. The direct path is typically referred to as the LoS component, and the scattered paths are referred to as the Non-LoS (NLoS) component. The interaction between these paths leads to the fading phenomena, which is often modeled statistically using *Rician* fading (sometimes written as Ricean fading). The main assumption is that the complex-valued channel coefficient between UE k and AP l in the complex baseband can be divided into two parts [53]:

$$h_{kl} = \overline{h}_{kl}e^{j\varphi_{kl}} + g_{kl} \tag{19}$$

where $\overline{h}_{kl} \geq 0$ is the magnitude of the LoS component between UE k and AP l and $\varphi_{kl} \in [0, 2\pi)$ is the corresponding phase-shift. The second part, g_{kl}, represents the NLoS component comprising all the scattered paths, of which each has roughly the same strength but is substantially weaker than the LoS component (this is why it needs to be modeled separately). Motivated by the central limit theorem, g_{kl} is modeled by a Gaussian distribution, which implies $g_{kl} \sim \mathcal{N}_{\mathbb{C}}(0, \beta_{kl})$, where $\beta_{kl} \geq 0$ is the variance. This is called Rayleigh fading since $|g_{kl}|$ is Rayleigh distributed, i.e.,

$|g_{kl}| \sim \text{Rayleigh}(\sqrt{\beta_{kl}/2})$. Under these assumptions, the magnitude $|h_{kl}|$ of the channel coefficient is Rice distributed, i.e., $|h_{kl}| \sim \text{Rice}(\overline{h}_{kl}, \sqrt{\beta_{kl}/2})$, which is why it is called Rician fading.

When the channel is assumed to be perfectly known at the receiver, the phase-shift φ_{kl} will not affect the communication performance since the receiver can compensate for it. Hence, it is common to omit φ_{kl} in the performance analysis of Rician fading channels. Consequently, h_{kl} in (19) can be drawn as $h_{kl} \sim \mathcal{N}_{\mathbb{C}}\left(h_{kl}, \beta_{kl}\right)$. However, we cannot neglect the phase φ_{kl} when analyzing practical systems where the receiver needs to estimate the channel, since the value of φ_{kl} varies at the same pace as g_{kl}, and φ_{kl} affects the strong LoS component where the impact of φ_{kl} cannot be ignored. However, the results obtained with a perfectly-known φ_{kl} can be interpreted as an upper bound on what is practically achievable.

Note that (19) represents the channel when single-antenna UEs and single-antenna APs are considered. In the case of single-antenna UEs and N-correlated-antenna APs, the channel between UE k and AP l is no longer a scalar but an N-dimensional vector as [54].

$$\mathbf{h}_{kl} = e^{j\varphi_{kl}}\mathbf{h}_{kl} + \mathbf{g}_{kl} \tag{20}$$

where $\mathbf{h}_{kl}$ and $\mathbf{g}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R}_{kl})$ represent the LoS and NLoS component respectively, and φ_{kl} is the common phase shift. If the phase-shift φ_{kl} is neglected, the channel $\mathbf{h}_{kl}$ can be viewed as a realization of the circularly symmetric complex Gaussian distribution [55,56].

$$\mathbf{h}_{kl} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{h}_{kl}, \mathbf{R}_{kl}\right) \tag{21}$$

Rayleigh fading is a tractable model for rich scattering scenarios without an LoS path, where the AP antenna array is surrounded by many scattering objects, as compared to the number of antennas per AP [57]. The Rayleigh fading channel is widely used to describe the basic properties of wireless propagation. Thus, the channel response h_{kl} is distributed as $h_{kl} \sim \mathcal{N}_{\mathbb{C}}(0, \beta_{kl})$, and the multi-antenna channel $\mathbf{h}_{kl}$ is distributed as

$$\mathbf{h}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R}_{kl}) \tag{22}$$

2.3. Duplex protocol

According to whether the uplink and downlink are separated in time or frequency, a CF mMIMO system can operate in the TDD or FDD mode. In the TDD mode, the signaling overhead scales with the number of the served UEs but is independent of the number of AP antennas due to the *channel reciprocity* that appears when transmitting in both directions in the same band. This means that one can perform the downlink precoding based on the CSI obtained from the uplink pilots since the channel response is the same in both directions. However, the channel reciprocity is not available in FDD since the uplink and downlink channels are in different bands, which introduces additional CSI acquisition and feedback overhead that is not only related to the number of served UEs but also the number of AP antennas. Most of the works on CF mMIMO systems assume the TDD mode to avoid the exchange of CSI and precoding/combining vectors [21,51,52,58].

However, in practice, this is not a design choice but rather determined by the spectrum license available for the system. Hence, it is also important to develop FDD-based CF mMIMO systems [59,60]. To limit the CSI acquisition and feedback overhead in FDD-based systems [59,60], exploit the property of the so-called *angle reciprocity*, which means the Angles-of-Departure (AoDs) are similar in both uplink and downlink. When the propagation channels are sufficiently sparse to utilize this property, the required overhead scales only with the number of the served UEs. In Ref. [59], the authors exploited the discrete Fourier transform operation and log-likelihood function to estimate the multipath component for the Angle-of-Arrival (AoA) and large-scale fading coefficients. Based on the estimated AoA, linear precoding/combining schemes were proposed with only scales with the number of the served UEs. In Ref. [60], the authors proposed a path to gain information feedback scheme where the required overhead for the channel vector quantization scales linearly with the number of dominating paths instead of the number of the serving antennas.

2.4. Scalability issues

Scalability is an essential issue for the network technology to be practically implemented, particularly when designing a technology where a large number of APs are supposed to cooperate. According to Ref. [21], a network is scalable if all the following tasks for the APs have finite complexity and resource requirement when the number of UEs tends to infinity:

1. Signal processing for channel estimation;
2. Signal processing for data reception and transmission;
3. Fronthaul signaling for data and CSI sharing;
4. Power control optimization.

The naive form of CF mMIMO in which each AP is required to process and share the data signals related to all UEs fails to be scalable since the computational complexity and fronthaul load associated with the above-listed tasks grow linearly (or faster) with the number of the UEs. The user-centric approach (also referred to as dynamic cooperating clustering [61,62]) takes the first step towards scalability by letting each AP only be responsible for a limited number of UEs in $\mathcal{D}_l \subset \{1, \ldots, K\}$, $l = 1, \ldots, L$, instead of all of them [20]. This makes a CF mMIMO system meet the first three conditions listed above if the cardinality $|\mathcal{D}_l|$ is constant as $K \to \infty$ for $l = 1, \ldots, L$. The reason comes from the fact that AP l only needs to compute the channel estimates and combining/precoding vectors for $|\mathcal{D}_l|$ UEs with a constant complexity as $K \to \infty$. Moreover, AP l only needs to receive/send the data related to these $|\mathcal{D}_l|$ UEs via the fronthaul network, which is a constant number as $K \to \infty$. How to select the UE sets $\mathcal{D}_l$ for $l = 1, \ldots, L$ in a scalable way while ensuring the service to all UEs is elaborated in Section 3.4. Suboptimal power control policies, such as fractional power control [63,64], are needed to limit the complexity of power control.

The signal processing procedure of the scalable CF mMIMO shares the similar methodology and mathematical expressions of the original alternative introduced earlier in this section by only letting the uplink combining vector $\mathbf{v}_{kl} = \mathbf{0}$ and the downlink precoding vector $\mathbf{w}_{kl} = \mathbf{0}$ for $k \notin \mathcal{D}_l$, $l = 1, \ldots, L$. The design of scalable combining and precoding vectors can be found in Section 3.3. Moreover, from the perspective of the UEs, we also denote by $\mathcal{M}_k \subset \{1, \ldots, L\}$, $k = 1, \ldots, K$, the subset of APs serving UE k.

2.5. Channel hardening and favorable propagation

Channel hardening and favorable propagation are the two basic virtues of cellular mMIMO [12, Sec. 2.5]. To be specific, *channel hardening* makes the fading channel between an AP and a UE behave as almost deterministic after the precoding/combining has been applied, when the number of serving antennas, i.e., LN grows large. Mathematically, this effect can be expressed as

$$\frac{\|\mathbf{h}_k\|^2}{\mathbb{E}\{\|\mathbf{h}_k\|^2\}} \to 1 \quad \text{as} \quad LN \to \infty \tag{23}$$

This is the ultimate form of spatial diversity, which removes the impact of the small-scale fading. In practice, a small amount of hardening is sufficient to alleviate the worst effects of fading and enable the resource allocation to be based on long-term statistics instead of the small-scale fading variations.

Beyond that, the directions of two UE channels are asymptotically orthogonal when the number of antennas approaches infinity, leading to the so-called *favorable propagation*, which is expressed as

$$\frac{\mathbf{h}_k^{\mathrm{H}}\mathbf{h}_i}{\sqrt{\mathbb{E}\{\|\mathbf{h}_k\|^2\}\mathbb{E}\{\|\mathbf{h}_i\|^2\}}} \to 0 \quad \text{when} \quad LN \to \infty,\ k \neq i \tag{24}$$

With approximate favorable propagation, one can get away with relatively simple signal processing techniques since interference vanishes automatically. However, some interference suppression is generally preferred.

3. Resource allocation and signal processing

Well-designed schemes and algorithms for resource allocation and signal processing are the keys to boosting the system performance of CF mMIMO networks. This section will provide a comprehensive survey of different categories of resource allocation and signal processing schemes, including the ones for channel estimation, combining and precoding, user access and association, and power control.

3.1. Channel estimation

The main benefits of serving a UE through multiple APs materialize when the APs have CSI so the received uplink signals can be coherently combined in the joint processing, and the downlink transmissions can be precoded to combine over the air coherently. Since the channels are time-varying, assuming that a complete and perfect CSI available at the APs and CPUs is not realistic. Consequently, developing accurate and resource-efficient channel estimation techniques is vital to good performance, especially the performance of legacy technologies. In this subsection, we provide an overview of the channel estimation techniques for CF mMIMO systems.

3.1.1. Normalized mean-squared error

Before we elaborate different estimators, we first introduce the *Mean-Squared Error (MSE)* which indicates the "distance" from the estimated channel and the actual channel, i.e., $\mathbb{E}\{\|\mathbf{h}-\hat{\mathbf{h}}\|^2\}$, where $\mathbf{h}$ and $\hat{\mathbf{h}}$ denote an arbitrary channel response and its estimate, respectively. However, the value of the MSE depends on the average channel gain, and hence a strong channel might have larger errors in absolute terms than a weaker one. To reasonably quantify the accuracy of an estimator, we consider the relative size of the error, i.e., the *Normalized MSE (NMSE)*. The NMSE between AP l and UE k using an arbitrary estimator is represented as

$$\mathrm{NMSE}_{kl} = \frac{\mathbb{E}\left\{\left\|\tilde{\mathbf{h}}_{kl}\right\|^2\right\}}{\mathbb{E}\{\|\mathbf{h}_{kl}\|^2\}} = \frac{\mathrm{tr}(\mathbf{C}_{kl})}{\mathrm{tr}(\mathbf{R}_{kl})} \tag{25}$$

where $\tilde{\mathbf{h}}_{kl} = \mathbf{h}_{kl} - \hat{\mathbf{h}}_{kl}$ and $\mathbf{C}_{kl} = \mathbb{E}\{\tilde{\mathbf{h}}_{kl}\tilde{\mathbf{h}}_{kl}^{\mathrm{H}}\}$ denote the estimation error of the considered estimator and its corresponding error correlation matrix, respectively.

When it comes to the collective channel of UE k, i.e., $\mathbf{h}_k$, the NMSE of its estimate $\hat{\mathbf{h}}_k$ can be computed as [4].

$$\mathrm{NMSE}_k = \frac{\mathbb{E}\left\{\left\|\mathbf{D}_k\tilde{\mathbf{h}}_k\right\|^2\right\}}{\mathbb{E}\{\|\mathbf{D}_k\mathbf{h}_k\|^2\}} = \frac{\sum_{l=1}^{L}\mathrm{tr}(\mathbf{D}_{kl}\mathbf{C}_{kl})}{\sum_{l=1}^{L}\mathrm{tr}(\mathbf{D}_{kl}\mathbf{R}_{kl})} \tag{26}$$

where $\tilde{\mathbf{h}}_k = [\tilde{\mathbf{h}}_{k1}^{\mathrm{T}}, \dots, \tilde{\mathbf{h}}_{kL}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{LN}$ and block-diagonal matrix $\mathbf{D}_k = \mathrm{diag}(\mathbf{D}_{k1}, \dots, \mathbf{D}_{kL})$ with $\mathbf{D}_{kl} = \mathbf{I}_N$ if $l \in \mathcal{M}_k$ and $\mathbf{D}_{kl} = \mathbf{0}_N$ otherwise. Note that (26) is not the sum or average of the individual NMSEs between UE K and its serving AP in $\mathcal{M}_k$, but contains a summation of MSEs in the numerator normalized by a summation of channel gains. Hence, it is the APs with strong channels that dominate in the summation.

3.1.2. Pilot-based channel estimation

The most commonly used approach for CSI acquisition is to transmit uplink pilot signals, where a predefined pilot signal is transmitted from the UE-side antenna, and all the antenna at the APs can simultaneously receive the transmission and compare it with the known pilot signal to estimate the channel response from the transmitting antenna. Suppose we instead need to estimate the channel response from two transmitting antennas. In that case, two orthogonal pilot signals are generally required to separate the signals from the two antennas (unless there is other prior information that allows for separation). The number of orthogonal pilot signals is proportional to the number of transmitting antennas, while any number of the receiving antennas can "listen" to the pilots simultaneously and estimate their respective channels to the transmitters. When the UEs transmit their pilots, the received signal can be represented as $\mathbf{y}_{t_k l}^{\mathrm{pilot}}$, which is given in (1), and based on that, there exist several different channel estimators.

Least-squares estimator If the statistics are unknown or unreliable, it might be necessary to consider the estimators that require no prior statistical information. The Least-Square (LS) estimator has been used for this purpose. The LS estimator minimizes $\|\mathbf{y}_{t_k l}^{\mathrm{pilot}} - \sqrt{p_k}\tau_p\hat{\mathbf{h}}_{kl}\|^2$, which is achieved by [12, Sec. 3.4]

$$\hat{\mathbf{h}}_{kl}^{\mathrm{LS}} = \frac{1}{\sqrt{\tau_p p_k}}\mathbf{y}_{t_k l}^{\mathrm{pilot}} = \mathbf{h}_{kl} + \sum_{i\in\mathcal{S}_k/\{k\}}\frac{\sqrt{p_i}}{\sqrt{p_k}}\mathbf{h}_{il} + \frac{1}{\sqrt{\tau_p p_k}}\mathbf{n}_{t_k l} \tag{27}$$

If the channels are Rayleigh fading as defined in (22), substituting (27) into (26), the NMSE of the LS estimator can be computed as

$$\mathrm{NMSE}_k^{\mathrm{LS}} = \frac{\sum_{l\in\mathcal{M}_k}\mathrm{tr}\left(\sum_{i\in\mathcal{S}_k/\{k\}}\frac{p_i}{p_k}\mathbf{R}_{il} + \frac{\sigma_{\mathrm{ul}}^2}{p_k\tau_p}\mathbf{I}_N\right)}{\sum_{l\in\mathcal{M}_k}\mathrm{tr}(\mathbf{R}_{kl})} \tag{28}$$

Note that the estimate $\hat{\mathbf{h}}_{kl}^{\mathrm{LS}}$ and the estimation error $\tilde{\mathbf{h}}_{kl}^{\mathrm{LS}}$ are correlated.

MMSE-type estimators The MMSE estimator has this name because it minimizes the MSE, and thereby also the NMSE. It takes different forms depending on the channel statistics. If the channels are Rayleigh fading, then the MMSE estimate $\hat{\mathbf{h}}_{kl}^{\mathrm{MMSE}}$ can be achieved by [12, Sec. 3.2]

$$\hat{\mathbf{h}}_{kl}^{\mathrm{MMSE}} = \sqrt{\tau_p p_k}\left(\mathbb{E}\left\{\mathbf{y}_{t_k l}^{\mathrm{pilot}}\left(\mathbf{y}_{t_k l}^{\mathrm{pilot}}\right)^{\mathrm{H}}\right\}\right)^{-1}\mathbf{y}_{t_k l}^{\mathrm{pilot}} \tag{29}$$

Substituting (29) into (26), the NMSE of the MMSE estimator can be computed as

$$\mathrm{NMSE}_k^{\mathrm{MMSE}} = \frac{\sum_{l\in\mathcal{M}_k}\mathrm{tr}(\mathbf{R}_{kl} - \tau_p p_k\mathbf{R}_{kl}\mathbf{\Psi}_{t_k l}^{-1}\mathbf{R}_{kl})}{\sum_{l\in\mathcal{M}_k}\mathrm{tr}(\mathbf{R}_{kl})} \tag{30}$$

This is the optimal channel estimator from an MSE and NMSE perspective, thus all other estimators described in this survey will provide larger MSEs. Recall that unlike the LS estimate, the MMSE estimate $\hat{\mathbf{h}}_{kl}^{\mathrm{MMSE}}$ and the estimation error $\tilde{\mathbf{h}}_{kl}^{\mathrm{MMSE}}$ are independent vectors.

Another alternative MMSE-type estimator is obtained by estimating each element of $\mathbf{h}_{kl}$ separately, thereby ignoring these correlations between the elements [56,65]. More precisely, we can consider one of the N elements in $\mathbf{y}_{t_k l}^{\mathrm{pilot}}$ at a time. The resulting Element-Wise MMSE (EW-MMSE) estimator that estimates the nth element $[\mathbf{h}_{kl}]_n$ is given as

$$\left[\hat{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}\right]_n = \frac{\sqrt{p_k\tau_p}[\mathbf{R}_{kl}]_{nn}}{\sum_{i\in\mathcal{S}_k}p_i\tau_p[\mathbf{R}_{il}]_{nn} + \sigma_{\mathrm{ul}}^2}\left[\mathbf{y}_{t_k l}^{\mathrm{pilot}}\right]_n \tag{31}$$

Since this is a type of MMSE estimator, the EW-MMSE estimate $[\hat{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}]_n$ and the corresponding estimation error $[\tilde{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}]_n$ are independent scalars. Substituting (31) into (26), the NMSE of the EW-MMSE estimator can be computed as

$$\mathrm{NMSE}_k^{\mathrm{EW-MMSE}} = \frac{\sum_{l\in\mathcal{M}_k}\sum_{n=1}^{N}[\mathbf{R}_{kl}]_{nn} - \frac{p_k\tau_p\left([\mathbf{R}_{kl}]_{nn}\right)^2}{\sum_{i\in\mathcal{S}_k} p_i\tau_p[\mathbf{R}_{il}]_{nn} + \sigma_{\mathrm{ul}}^2}}{\sum_{l\in\mathcal{M}_k}\mathrm{tr}(\mathbf{R}_{kl})} \tag{32}$$

However, there can be cross-correlation between the estimate and the estimation error for different antennas, and its existence demonstrates the suboptimality of EW-MMSE. An optimal estimator exploits all correlation to lower the MSE. The EW-MMSE estimator results in larger estimation errors since the correlation between the variables are not utilized to improve the estimation quality. One benefit of the EW-MMSE estimator is that it requires less statistical information since only the diagonals of the spatial correlation matrices are utilized. Moreover, it has lower computational complexity than the MMSE estimator, except in the particular case where all the spatial correlation matrices are diagonal so that one can estimate each channel element separately without a performance loss.

Since MMSE-type estimators depend on the channel statistics, they also take different forms when the channel model is changed. When considering Rician fading channels, as defined in (20), there are different estimators of the MMSE-type. If the LoS component $\mathbf{h}_{kl}$, the spatial correlation matrix $\mathbf{R}_{kl}$, and the phase-shift φ_{kl} are known, the Phase-Aware MMSE (PA-MMSE) estimate of $\mathbf{h}_{kl}$ is given as [53,54].

$$\hat{\mathbf{h}}_{kl}^{\mathrm{PA-MMSE}} = \mathbf{h}_{kl}e^{j\varphi_{kl}} + \sqrt{\tau_p p_k}\mathbf{R}_{kl}\mathbf{\Psi}_{t_k l}^{-1}(\mathbf{y}_{t_k l}^{\mathrm{pilot}} - \mathbf{z}_{t_k l}) \tag{33}$$

where $\mathbf{z}_{t_k l} = \sum_{i\in\mathcal{S}_k}\sqrt{p_i\tau_p}\mathbf{h}_{il}e^{j\varphi_{il}}$. The estimate $\hat{\mathbf{h}}_{kl}^{\mathrm{PA-MMSE}}$ and estimation error $\tilde{\mathbf{h}}_{kl}^{\mathrm{PA-MMSE}}$ are independent random variables.

If the channel statistics $\mathbf{h}_{kl}$, $\mathbf{R}_{kl}$ are available while the phase φ_{kl} is unknown and uniformly distributed from 0 to 2π, then the Linear MMSE (LMMSE) estimator of $\mathbf{h}_{kl}$ is [53].

$$\hat{\mathbf{h}}_{kl}^{\mathrm{LMMSE}} = \sqrt{p_k}\mathbf{R}_{kl}^{'}\left(\mathbf{\Psi}_{t_k l}^{'}\right)^{-1}\mathbf{y}_{t_k l}^{\mathrm{pilot}} \tag{34}$$

where $\mathbf{R}'_{kl} = \mathbf{R}_{kl} + \mathbf{h}_{kl}\mathbf{h}_{kl}^H$ and $\mathbf{\Psi}'_{t_k l} = \sum_{i\in\mathcal{S}_k} p_i\tau_p\left(\mathbf{R}_{il} + \mathbf{h}_{il}\mathbf{h}_{il}^H\right) + \sigma_{\mathrm{ul}}^2\mathbf{I}_N$. The LMMSE estimation $\hat{\mathbf{h}}_{kl}^{\mathrm{LMMSE}}$ and the estimation error $\tilde{\mathbf{h}}_{kl}^{\mathrm{LMMSE}}$ are independent random variables.

It is worth noting that the aforementioned pilot-based schemes have same names as the ones used in the cellular mMIMO literature, because they are derived to minimize the same general metrics (see Table 3). However, the estimator expressions differ since pilot contamination affects the systems differently, and the notation is different. We also notice that the CF mMIMO estimators can be computed separately at each AP (instead of centrally at one CPU as in cellular mMIMO) since the channel vectors are independent between APs. Therefore, the estimation schemes are specially tailored for CF mMIMO systems. A comparison between the aforementioned estimators over spatially correlated Rayleigh fading channels is provided in Fig. 4, where the average NMSE of $K = 50$ UEs is shown as a decreasing function of the uplink transmit power per UE, p, with $L = 100$ APs and each equipped with $N = 4$ antennas. It can be seen that the MMSE-type estimators significantly outperform the LS estimator since the LS estimator has no prior information about the channel statistics. As mentioned before, the EW-MMSE estimator results in a larger average NMSE compared to the MMSE estimator since the former only exploits partial statistical information while the latter relies on the full statistical channel knowledge. The performance gap between the MMSE and EW-MMSE estimators will substantially decrease when considering the spatially correlated Rician fading channels since the existence of LoS components weakens the effect of spatial correlation, and the diagonals of the correlation matrix of these two MMSE-type estimators are identical. Moreover, all aforementioned channel estimation schemes are summarized in Table 4, where the abbreviation "In." stands for whether the channel estimate and the estimation error are independent, and the closed-form expressions can be found in the references.

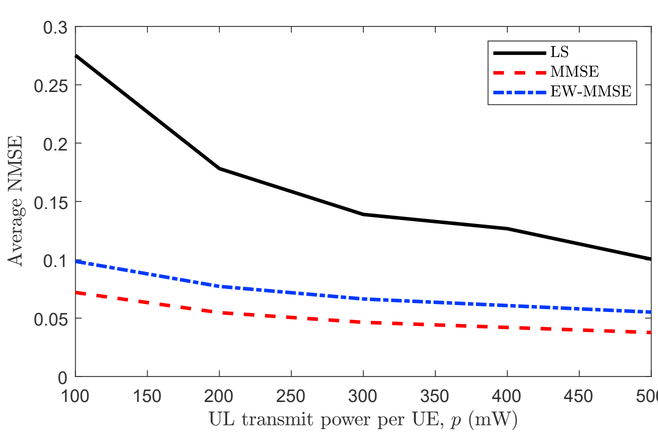


Fig. 4. Average NMSE versus p with MMSE [12], EW-MMSE [56,65], and LS [12] estimator.

Table 3
Important mathematical symbols.

Notation	Definition
K, L, N	Number of the UEs, APs,and antennas per AP
τ_p	Number of orthogonal pilot sequences
k, i	Index of the UEs
l, j	Index of the APs
t_k	Index of the pilot assigned to UE k
$\mathcal{M}_k$	Subset of APs serving UE k
$\mathcal{D}_l$	Subset of UEs served by AP l
$\mathcal{S}_k$	Subset of UEs sharing pilot t_k,including UE k
$\mathcal{P}_k$	Subset of UEs served by partially the same APs as UE k, including UE k
$\mathbf{h}_{kl}$	Channel response between UE k and AP l
$\hat{\mathbf{h}}_{kl}$	Channel estimate of $\mathbf{h}_{kl}$
β_{kl}	Large-scale fading coefficient of UE k and AP l
$\mathbf{v}_{kl}$	Combining vector that AP l selects for UE k
$\mathbf{w}_{kl}$	Precoding vector that AP l selects for UE k

3.1.3. ML-based channel estimation

Machine Learning (ML) is a powerful tool for identifying the structure and data and making decisions, and thus can particularly be utilized for reducing the computational complexity of known algorithms or learning mappings between the known and unknown variables [66]. Although the MMSE estimator is optimal for Rayleigh and Rician fading channels, the price to pay is the high computational complexity from inverting a matrix (as shown in (29)), if each AP has many antennas. This motivates the design of the ML-based channel estimation which might reduce the online computational complexity by exploiting data-driven signal processing algorithms. Besides, practical wireless channels for future wireless communications (e.g., mmWave channels) can only be approximately described by Rayleigh and Rician fading, thus the MMSE-type estimators are not optimal in practice. Hence, an ML-based estimator can potentially improve the estimation quality. In Ref. [67], the authors proposed a Fast and Flexible Denoising Convolutional Neural Network (FFDNet) for the channel estimation in CF mMIMO systems. By introducing a noise level map as input sub-images, a single neural network can be used to handle different noise levels and reduce the waiting time for training and testing. The Convolutional Blind Denoising Network (CBDNet) is developed to improve the blind denoising performance for real-world noisy images, thus it can be used for boosting the quality of channel estimation by regarding the channel matrix as an image. The CBDNet-based channel estimation for mmWave mMIMO systems was proposed in Ref. [68], where the sparsity feature of the mmWave channel was exploited to achieve a notable performance gain with a wide range of SNRs and fast convergence.

Fig. 5 compares the NMSE performance of CBDNet- and FFDNet-

Table 4
Pilot training-based channel estimation.

Scheme	Estimate of $\mathbf{h}_{kl}$	Mean and Covariance of Estimate	Mean and Covariance of Estimation Error	In.
LS	$\hat{\mathbf{h}}_{kl}^{\mathrm{LS}} = \frac{1}{\sqrt{p_k\tau_p}}\mathbf{y}_{t_k l}^{\mathrm{pilot}}$	$\mathbb{E}\{\hat{\mathbf{h}}_{kl}^{\mathrm{LS}}\} = \mathbf{0}$ $\mathbb{C}\{\hat{\mathbf{h}}_{kl}^{\mathrm{LS}}\} = \sum_{i\in\mathcal{S}_k}\frac{p_i}{p_k}\mathbf{R}_{il} + \frac{\sigma_{\mathrm{ul}}^2}{p_k\tau_p}\mathbf{I}_N$	$\mathbb{E}\{\tilde{\mathbf{h}}_{kl}^{\mathrm{LS}}\} = \mathbf{0}$ $\mathbb{C}\{\tilde{\mathbf{h}}_{kl}^{\mathrm{LS}}\} = \mathbb{C}\{\hat{\mathbf{h}}_{kl}^{\mathrm{LS}}\} - \mathbf{R}_{kl}$	×
MMSE	$\hat{\mathbf{h}}_{kl}^{\mathrm{MMSE}} = \sqrt{\tau_p p_k}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{y}_{t_k l}^{\mathrm{pilot}}$	$\mathbb{E}\{\hat{\mathbf{h}}_{kl}^{\mathrm{MMSE}}\} = \mathbf{0}$ $\mathbb{C}\{\hat{\mathbf{h}}_{kl}^{\mathrm{MMSE}}\} = \tau_p p_k\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{R}_{kl}$	$\mathbb{E}\{\tilde{\mathbf{h}}_{kl}^{\mathrm{MMSE}}\} = \mathbf{0}$ $\mathbb{C}\{\tilde{\mathbf{h}}_{kl}^{\mathrm{MMSE}}\} = \mathbf{R}_{kl} - \mathbb{C}\{\hat{\mathbf{h}}_{kl}^{\mathrm{MMSE}}\}$	✓
EW-MMSE	$[\hat{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}]_n = \frac{\sqrt{p_k\tau_p}[\mathbf{R}_{kl}]_{nn}}{\sum_{i\in\mathcal{S}_k}p_i\tau_p[\mathbf{R}_{il}]_{nn} + \sigma_{\mathrm{ul}}^2} \times [\mathbf{y}_{t_k l}^{\mathrm{pilot}}]_n$	$\mathbb{E}\{[\hat{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}]_n\} = 0$ $\mathbb{C}\{[\hat{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}]_n\} = \frac{p_k\tau_p\left([\mathbf{R}_{kl}]_{nn}\right)^2}{\sum\limits_{i\in\mathcal{S}_k}p_i\tau_p[\mathbf{R}_{il}]_{nn} + \sigma_{\mathrm{ul}}^2}$	$\mathbb{E}\{[\tilde{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}]_n\} = 0$ $\mathbb{C}\{[\tilde{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}]_n\} = [\mathbf{R}_{kl}]_{nn} - \mathbb{C}\{[\hat{\mathbf{h}}_{kl}^{\mathrm{EW-MMSE}}]_n\}$	✓
PA-MMSE	$\hat{\mathbf{h}}_{kl}^{\mathrm{PA-MMSE}} = \mathbf{h}_{kl}e^{j\varphi_{kl}} + \sqrt{p_k}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_k l}^{-1}(\mathbf{y}_{t_k l}^{\mathrm{pilot}} - \mathbf{z}_{t_k l})$	$\mathbb{E}\{\hat{\mathbf{h}}_{kl}^{\mathrm{PA-MMSE}}\|\varphi_{kl}\} = \mathbf{h}_{kl}e^{j\varphi_{kl}}$ $\mathbb{C}\{\hat{\mathbf{h}}_{kl}^{\mathrm{PA-MMSE}}\|\varphi_{kl}\} = p_k\tau_p\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{R}_{kl}$	$\mathbb{E}\{\tilde{\mathbf{h}}_{kl}^{\mathrm{PA-MMSE}}\} = \mathbf{0}$ $\mathbb{C}\{\tilde{\mathbf{h}}_{kl}^{\mathrm{PA-MMSE}}\} = \mathbf{R}_{kl} - p_k\tau_p\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{R}_{kl}$	✓
LMMSE	$\hat{\mathbf{h}}_{kl}^{\mathrm{LMMSE}} = \sqrt{p_k}\mathbf{R}_{kl}^{'}\left(\boldsymbol{\Psi}_{t_k l}^{'}\right)^{-1}\mathbf{y}_{t_k l}^{\mathrm{pilot}}$	$\mathbb{E}\{\hat{\mathbf{h}}_{kl}^{\mathrm{LMMSE}}\} = \mathbf{0}$ $\mathbb{C}\{\hat{\mathbf{h}}_{kl}^{\mathrm{LMMSE}}\} = p_k\tau_p\mathbf{R}_{kl}^{'}\left(\boldsymbol{\Psi}_{t_k l}^{'}\right)^{-1}\mathbf{R}_{kl}^{'}$	$\mathbb{E}\{\tilde{\mathbf{h}}_{kl}^{\mathrm{LMMSE}}\} = \mathbf{0}$ $\mathbb{C}\{\tilde{\mathbf{h}}_{kl}^{\mathrm{LMMSE}}\} = \mathbf{R}_{kl}^{'} - \mathbb{C}\{\hat{\mathbf{h}}_{kl}^{\mathrm{LMMSE}}\}$	✓

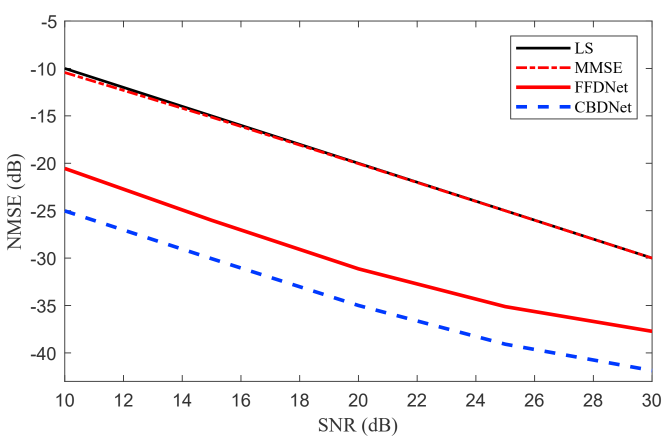


Fig. 5. NMSE versus SNR with CBDNet [68], FFDNet [67], LS [12], and MMSE [12] estimator.

based channel estimators with the conventional LS and MMSE estimators in a mmWave channel setup. We first use the fast Fourier transform-based modulation and demodulation schemes to expose the low-sparse information in the angle and delay domains. To be specific, the ML-based methods learn the sparse virtual-channel coefficients information of the mmWave channel, with $\leq L$ scattering clusters having non-zero coefficients, and introduce the residual network to capture the low-dimensional sparse channel subspace that carries most of the power and propagation angle information efficiently. It can be seen that the ML-based methods (i.e., FFDNet and CBDNet) significantly outperform the conventional methods (i.e., LS and MMSE) since the ML-based methods learn certain information about the propagation environment while the counterparts only assume the covariance information that cannot fully capture the sparsity. It is worth noting that the more prior information that one has about the propagation environment, the better the estimates can become. In theory, one can develop a new MMSE estimator for the scenario simulated here, which will be optimal in terms of NMSE if the same side information learned by ML-based methods is available. However, this prior information may be hard to describe mathematically and must be extracted from the data; that is, when the ML approaches (e.g., FFDNet and CBDNet) prove their abilities. Moreover, it is illustrated that FFDNet has a larger NMSE than CBDNet since FFDNet only fits narrow noise levels and lacks the adaptability to out-of-trained range noise levels. While CBDNet can deal with the whole and beyond training SNR range using continuous nonlinear joint loss function to enlarge the SNR range and achieve fast convergence because CBDNet converts the loss to the same order of magnitude.

3.1.4. Other methods

Beyond the pilot- and ML-based channel estimation methods, other options exist for their specific applications in CF mMIMO systems. Recall that pilot contamination is an inevitable phenomenon introduced by pilot reuse, which can substantially deteriorate the estimation accuracy and the system performance. It might be less of an issue in CF mMIMO than in the conventional mMIMO due to the distributed nature, where each AP has relatively few antennas, but it cannot be neglected. To mitigate pilot contamination [69], proposed a Time-Of-Arrival (TOA)-based scheme that first estimates the TOA of the multipath channel and then filters out interfering signals of paths originating from distant APs. It is shown that the proposed TOA-based scheme could outperform the LS estimator, given the TOA estimation is accurate. The TOA-based channel estimation only depends on the currently received signal without the need for channel statistics, which is the key advantage of this scheme that makes it more practically applicable.

In CF mMIMO systems, additional fronthaul overhead is required to exchange CSI between the APs and CPU, making blind channel estimation schemes attractive from the overhead perspective. Independent Component Analysis (ICA), which is one of the blind source separation approaches, was employed in Ref. [70] for blind channel estimation and signal detection in CF mMIMO systems. The proposed scheme used ICA to separate and decode the received signals and to estimate the channels. The estimated channel energy was used to differentiate the in-cell signals and the neighboring cell signals. The reference bits were applied to identify the desired signal among signals within a cell. However, this ICA-based method suffers from an error floor at high SINR due to inadequate ambiguity elimination.

When considering FDD mode, the key challenges are mainly CSI acquisition and feedback overhead. Even if the channel reciprocity does not hold, we can exploit the angle reciprocity, in which the AoDs are similar in both the uplink and downlink. In Ref. [71], the authors proposed a Discrete Fourier Transform (DFT)-based channel estimation scheme in an FDD CF mMIMO system where the angle reciprocity of multipath components in both the uplink and downlink is exploited. Based on the DFT operation and the log-likelihood function where the angle rotation is with a tiny amount of training overhead, the required CSI requisition overhead scales only with the number of served UEs.

3.2. Receive combining

In the uplink, the APs utilize the CSI that is acquired by channel estimation to perform the coherent signal processing, which is called *receive combining*. Receive combining is a linear projection that transforms the vector channels into effective scalar channels that support higher SEs than in the case where only one single-antenna AP serves each UE [72]. The purpose of the receive combining is to make the desired signal much stronger than the sum of interfering signals, and noise [73–75], which requires CSI. Different combining methods lead to substantially different SEs and computational complexities. In this subsection, we comprehensively introduce the receive combining used in CF mMIMO networks with a centralized or distributed operation.

3.2.1. MR combining

The scheme with the lowest complexity is Maximum Ratio (MR) combining, defined as [12, Sec. 4.1]

$$\mathbf{v}_{kl}^{\mathrm{MR}} = \hat{\mathbf{h}}_{kl} \tag{35}$$

This is a vector that maximizes the ratio $|\mathbf{v}_{kl}^H\hat{\mathbf{h}}_{kl}|^2/||\mathbf{v}_{kl}||^2$ between the power of the desired signal and the squared norm of the combining vector [72]. It shows that MR coherently combines all the received energy from the desired signal because the combining vector is matched to the channel response of the desired UE [76–78].

3.2.2. MMSE-like combining

Although MR maximizes the gain of the desired signal, it might not be the preferable choice when there are interfering signals. The MMSE-like methods can be utilized to identify a tradeoff between maximizing the signal gain and rejecting interference. There are different methods for distributed and centralized operation. When the centralized combining is considered, the Centralized MMSE (C-MMSE) combining is given as [12, Sec. 4.1]

$$\mathbf{v}_{k}^{\mathrm{C-MMSE}} = p_k\left(\sum\nolimits_{i=1}^{K} p_i\hat{\mathbf{h}}_i\hat{\mathbf{h}}_i^H + \sum\nolimits_{i=1}^{K} p_i\mathbf{C}_i + \sigma_{\mathrm{ul}}^2\mathbf{I}\right)^{-1}\hat{\mathbf{h}}_k \tag{36}$$

and minimizes the conditional MSE $\mathbb{E}\{|s_k - \mathbf{v}_k^H\mathbf{y}^{\mathrm{ul}}|^2|\{\hat{\mathbf{h}}_i\}\}$ between the desired signal and the centralized receive combined signal, where the expectation is computed conditioned on the centralized channel estimates. The C-MMSE combining can also be shown to maximize the SINR of UE k [49,50]. It has a relatively high computational complexity but since the computation in (36) is performed at the CPU, which is generally assumed to have high computational capability, it can be practically implementable. To reduce the complexity and enable the distributed implementation, we will consider several MMSE-like combining methods that are specially developed for CF mMIMO. As a distributed combining scheme directly inspired by C-MMSE, AP l can use local MMSE (L-MMSE) combining [49].

$$\mathbf{v}_{kl}^{\mathrm{L-MMSE}} = p_k\left(\sum\nolimits_{i=1}^{K} p_i\hat{\mathbf{h}}_{il}\hat{\mathbf{h}}_{il}^H + \sum\nolimits_{i=1}^{K} p_i\mathbf{C}_{il} + \sigma_{\mathrm{ul}}^2\mathbf{I}\right)^{-1}\hat{\mathbf{h}}_{kl} \tag{37}$$

when detecting the data from UE k. This combining scheme has received its name from the fact that it minimizes the conditional MSE $\mathbb{E}\{|s_k - \mathbf{v}_{kl}^H\mathbf{y}_l^{\mathrm{ul}}|^2|\{\hat{\mathbf{h}}_{il}\}\}$ between the desired signal s_k and the local receive combined signal $\mathbf{v}_{kl}^H\mathbf{y}_l^{\mathrm{ul}}$ at AP l, where the expectation is computed conditioned on the local channel estimates.

It can be observed from (37) and (36) that we need to compute all the K MMSE channel estimates $\{\hat{\mathbf{h}}_{il} : i = 1, \ldots, K\}$ at any AP l that is serving UE k. Therefore, the total number of complex multiplications required by L-MMSE and C-MMSE combining schemes, unfortunately, grow with K, thus making the complexity unscalable. To solve this issue, the alternative Partial MMSE (P-MMSE) and Local P-MMSE (LP-MMSE) are proposed in Ref. [21]. The main idea of P-MMSE and LP-MMSE is that the interference that affects UE k is mainly generated by a small subset of other UEs [50]. Therefore, only the UEs that are partially served by the same APs as UE k should be included in the expression in (37) and (36). These UEs have indices in the set

$$\mathcal{P}_k = \{i : \mathcal{M}_k \cap \mathcal{M}_i \neq \varnothing\} \tag{38}$$

By utilizing $\mathcal{P}_k$, the P-MMSE and LP-MMSE combining vectors are given as

$$\begin{aligned}\mathbf{v}_{k}^{\mathrm{P-MMSE}} &= p_k\left(\sum_{i\in\mathcal{P}_k} p_i\mathbf{D}_k\hat{\mathbf{h}}_i\hat{\mathbf{h}}_i^{\mathrm{H}}\mathbf{D}_k\right.\\ &\quad\left. + \mathbf{D}_k\left(\sum_{i\in\mathcal{P}_k} p_i\mathbf{C}_i + \sigma_{\mathrm{ul}}^2\mathbf{I}_{LN}\right)\mathbf{D}_k,\right)^{\dagger}\mathbf{D}_k\hat{\mathbf{h}}_k\end{aligned} \tag{39}$$

$$\mathbf{v}_{k}^{\mathrm{LP-MMSE}} = p_k\left(\sum_{i\in\mathcal{P}_k} p_i\left(\hat{\mathbf{h}}_{il}\hat{\mathbf{h}}_{il}^H + \mathbf{C}_{il}\right) + \sigma_{\mathrm{ul}}^2\mathbf{I}_{LN}\right)^{-1}\hat{\mathbf{h}}_{kl}. \tag{40}$$

The structure of the MMSE-like combining is quite intuitive. Let us take L-MMSE as an example. The matrix that is inverted in (37) is the conditional correlation matrix $\mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}} = \mathbb{E}\{\mathbf{y}_l^{\mathrm{ul}}(\mathbf{y}_l^{\mathrm{ul}})^{\mathrm{H}}|\{\hat{\mathbf{h}}_{il}\}\}$ of the received signal, given the current set of channel estimates. The multiplication $\mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}}^{-1/2}\mathbf{y}_l^{\mathrm{ul}}$ corresponds to the whitening of the received signal; that is, $\mathbb{E}\{\mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}}^{-1/2}\mathbf{y}_l^{\mathrm{ul}}(\mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}}^{-1/2}\mathbf{y}_l^{\mathrm{ul}})^{\mathrm{H}}|\{\hat{\mathbf{h}}_{il}\}\} = \mathbf{I}$. If we denote the whitened combining vector as $\mathbf{u}_{kl}$, it is related to the original combining vector as $\mathbf{v}_{kl} = \mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}}^{-1/2}\mathbf{u}_{kl}$. The highest desired signal power is now received from the spatial direction $\mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}}^{-1/2}\hat{\mathbf{h}}_{kl}$ and due to the whitening, which makes the total power equal in all directions, the interference plus noise power is the lowest in this direction. Hence, the optimal whitened combining vector can be selected as $\mathbf{u}_{kl} = \mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}}^{-1/2}\hat{\mathbf{h}}_{kl}$. This results into $\mathbf{v}_{kl} = \mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}}^{-1/2}\mathbf{u}_{kl} = \mathbf{C}_{\mathbf{y}_l^{\mathrm{ul}}}^{-1}\hat{\mathbf{h}}_{kl}$. Therefore, the MMSE-like combining is obtained by whitening followed by the MR combining [12].

Although the MMSE-like combining is optimal, the research literature contains other schemes as well. There are two main reasons for that. Firstly, the C-MMSE scheme has high complexity since there is an $LN \times LN$ matrix inverse in (36). Secondly, the performance of MMSE-like schemes is hard to analyze mathematically, while there are alternative schemes that can give more insightful closed-form SE expressions.

3.2.3. ZF-like combining

In the centralized operation, if the channel conditions are good, we can neglect all the correlation matrices in (36) and obtain the Regularized Zero-Forcing (RZF) combining vector [72,76].

$$\mathbf{v}_k^{\mathrm{RZF}} = \hat{\mathbf{H}}\left(\hat{\mathbf{H}}^{\mathrm{H}}\hat{\mathbf{H}} + \sigma_{\mathrm{ul}}^2\mathbf{P}^{-1}\right)^{-1}\hat{\mathbf{e}}_k \tag{41}$$

where $\hat{\mathbf{H}} = \left[\hat{\mathbf{H}}_1^{\mathrm{UE}}, \ldots, \hat{\mathbf{H}}_K^{\mathrm{UE}}\right] \in \mathbb{C}^{LN\times K}$, $\hat{\mathbf{H}}_k^{\mathrm{UE}} = \left[\hat{\mathbf{h}}_{1k}^T, \cdots \hat{\mathbf{h}}_{Lk}^T\right]^T \in \mathbb{C}^{LN}$ for

$k \in \{1, \ldots, K\}$, $\mathbf{P} = \mathrm{diag}(p_1, \ldots, p_K) \in \mathbb{C}^{K\times K}$, and $\hat{\mathbf{e}}_k$ is the kth column of $\mathbf{I}_K$. The RZF scheme can also be realized in a distributed fashion, which called Local RZF (L-RZF) scheme

$$\mathbf{v}_{kl}^{\mathrm{L-RZF}} = \hat{\mathbf{H}}_l(\hat{\mathbf{H}}_l^{\mathrm{H}}\hat{\mathbf{H}}_l + \sigma_{\mathrm{ul}}^2\mathbf{P}^{-1})^{-1}\hat{\mathbf{e}}_k \tag{42}$$

where $\hat{\mathbf{H}}_l = [\hat{\mathbf{h}}_{1l}, \ldots, \hat{\mathbf{h}}_{Kl}]$. When the SNR is high, the combining expression in (41) can be further approximated as

$$\mathbf{v}_k^{\mathrm{ZF}} = \hat{\mathbf{H}}(\hat{\mathbf{H}}^{\mathrm{H}}\hat{\mathbf{H}})^{-1}\hat{\mathbf{e}}_k \tag{43}$$

under the name of centralized Zero-Forcing (ZF) combining.

Unlike the centralized ZF combining, the Full-Pilot ZF (FZF) combining can suppress interference in a fully distributed, coordinated, and scalable fashion [79,80]. Besides, the computation of FZF combining has much lower complexity than the centralized ZF. When considering mutually orthogonal pilot sequences and uncorrelated Rayleigh channels between APs and UEs, the channel estimate of $\mathbf{h}_{kl}$ is given as

$$\hat{\mathbf{h}}_{kl} = c_{kl}\left(\sum_{k=1}^{K}\sqrt{p_k}\,\mathbf{h}_{kl}\boldsymbol{\phi}_{t_k}^{H} + \mathbf{N}_l\right)\boldsymbol{\phi}_{t_k} \tag{44}$$

where

$$c_{kl} \triangleq \frac{\sqrt{p_k\tau_p}\beta_{kl}}{\tau_p\sum_{i\in\mathcal{S}_k}p_i\beta_{kl} + \sigma^2} \tag{45}$$

$\boldsymbol{\phi}_{t_k}$ is the τ_p-length pilot signals assigned to UE k, and $\mathbf{N}_l \in \mathbb{C}^{N\times\tau_p}$ is a Gaussian noise matrix with i.i.d. $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ elements. The local combining vector that AP l selects for UE k, $\mathbf{v}_{t_k l}^{\mathrm{FZF}} \in \mathbb{C}^{N\times 1}$ is given by

$$\mathbf{v}_{t_k l}^{\mathrm{FZF}} = c_{kl}\mathbf{H}_l(\mathbf{H}_l^{\mathrm{H}}\mathbf{H}_l)^{-1}\mathbf{e}_{t_k} \tag{46}$$

where

$$\overline{\mathbf{H}}_l = \sum_{k=1}^{K}\sqrt{p_k}\mathbf{h}_{kl}\boldsymbol{\phi}_{t_k}^{\mathrm{H}}\boldsymbol{\Phi} + \mathbf{N}_l\boldsymbol{\Phi} \tag{47}$$

which is connected to the respective channel estimate by

$$\hat{\mathbf{h}}_{kl} = c_{kl}\mathbf{H}_l\mathbf{e}_{t_k} \tag{48}$$

where $\boldsymbol{\Phi} = \left[\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{\tau_p}\right] \in \mathbb{C}^{\tau_p\times\tau_p}$ and $\mathbf{e}_{t_k}$ denotes the t_kth column of $\mathbf{I}_{\tau_p}$.

Employing the Partial FZF (P-FZF) combining leads to more array gains with the cost of only suppressing the partial interference. Specifically, AP l employs the P-FZF combining for strong UEs whose channel gains are large $\mathcal{T}_l \subset \{1, \ldots, K\}$ and the MR combining for UEs with a poor channel condition $\mathcal{E}_l \subset \{1, \ldots, K\}$. Since only strong UEs use the P-FZF combining, we define $\tau_{\mathcal{T}_l}$ as the number of different pilots used by the UEs $\in \mathcal{T}_l$ and $\mathcal{R}_{\mathcal{T}_l} = (r_{l,1}, \ldots, r_{l,\tau_{\mathcal{T}_l}})$ as the set of the corresponding pilot indices. Therefore, the pilot-book matrix for UEs $\in \mathcal{T}_l$ is given by $\boldsymbol{\Phi}_{\mathcal{T}_l} = \boldsymbol{\Phi}\mathbf{E}_{\mathcal{T}_l}$, where $\mathbf{E}_{\mathcal{T}_l} = (\mathbf{e}_{r_{l,1}}, \ldots, \mathbf{e}_{r_{l,\tau_{\mathcal{T}_l}}}) \in \mathbb{C}^{\tau_p\times\tau_{\mathcal{T}_l}}$, and $\mathbf{e}_{r_{l,i}}$ is the $r_{l,i}$th column of $\mathbf{I}_{\tau_p}$. With respect to $\boldsymbol{\Phi}_{\mathcal{T}_l}$, we define $j_{kl} \in \{1, \ldots, \tau_{\mathcal{T}_l}\}$ the index. Let $\epsilon_{j_{kl}} \in \mathbb{C}^{\tau_{\mathcal{T}_l}\times 1}$ be the j_{kl}th column of $\mathbf{I}_{\tau_{\mathcal{T}_l}}$, and it leads to $\mathbf{E}_{\mathcal{T}_l}\epsilon_{j_{kl}} = \mathbf{e}_{t_k}$. Then, the P-FZF combining for UE $k \in \mathcal{T}_l$ at AP l is given as

$$\mathbf{v}_{t_k l}^{\mathrm{P-FZF}} = c_{kl}\mathbf{H}_l\mathbf{E}_{\mathcal{T}_l}(\mathbf{E}_{\mathcal{T}_l}^{\mathrm{H}}\mathbf{H}_l^{\mathrm{H}}\mathbf{H}_l\mathbf{E}_{\mathcal{T}_l})^{-1}\epsilon_{j_{kl}} \tag{49}$$

To improve the service quality of the weak UEs, which is the main advantage of CF mMIMO compared with cellular systems, we can alternatively apply the Protective Weak P-FZF (PWP-FZF) combining for weak UEs to significantly reduce the intra-group interference [81]. The main idea of PWP-FZF is to force the MR combining vector to take place in the orthogonal complement of $\mathbf{H}_l\mathbf{E}_{\mathcal{T}_l}$, which is the effective channels of UEs in $\mathcal{T}_l$. With PWP-FZF, the MR combining used at AP l for UEs in $\mathcal{E}_l$ is now given by

$$\mathbf{v}_{kl}^{\mathrm{PMR}} = c_{kl}\mathbf{J}_l\mathbf{H}_l\mathbf{e}_{t_k} \tag{50}$$

where

$$\mathbf{J}_l = \mathbf{I}_N - \mathbf{H}_l\mathbf{E}_{\mathcal{T}_l}\left(\mathbf{E}_{\mathcal{T}_l}^{\mathrm{H}}\mathbf{H}_l^{\mathrm{H}}\mathbf{H}_l\mathbf{E}_{\mathcal{T}_l}\right)^{-1}\mathbf{E}_{\mathcal{T}_l}^{\mathrm{H}}\mathbf{H}_l^{\mathrm{H}} \tag{51}$$

represents the projection matrix onto the orthogonal complement of $\mathbf{H}_l\mathbf{E}_{\mathcal{T}_l}$.

3.2.4. MMSE-SIC combining

All the combining schemes mentioned above are based on the linear receive combining. Still, another benefit of centralizing the signal processing at the CPU is that more advanced decoding methods can be used since system-wide CSI and high computational resources are available. The potential benefits of the MMSE-based non-linear Successive Interference Cancelation (SIC) method are investigated in Ref. [49], which means that the CPU decodes one UE signal at a time, and then sequentially subtracts interference that the decoded signal causes to the remaining signals. However, the numerical results in Ref. [49] show that the non-linear MMSE-SIC can only achieve a minor gain over the linear MMSE in terms of the average SE in the uplink when the favorable propagation phenomenon exists. The reason is that the SIC method is only effective when there are a few strongly interfering UEs, while CF mMIMO is more characterized by having many UEs that cause little interference to each other.

In Fig. 6, the Cumulative Distribution Function (CDF) of the uplink SE per UE is shown for the MR, FZF, P-FZF, PWP-FZF, L-RZF, and L-MMSE combining schemes with $L = 25$, $K = 10$, $N = 8$, $\tau_p = 7$ and $p_k = 100$ mW for each UE when using LSFD. The performance gap between the MR combining and the ZF-based schemes is quite significant, especially for UEs with large channel gains. It results from the impact of the inter-user interference while FZF, P-FZF, and PWP-FZF combining schemes all can suppress that interference. Besides, the advantage of employing P-FZF and PWP-FZF rather than FZF is noticeable. FZF spends τ_p degrees of freedom to cancel the pilot contamination and inter-user interference while P-FZF and PWP-FZF only spend $\tau_{\mathcal{T}_l}$ degrees of freedom and take advantage of a larger array gain. Compared with P-FZF, PWP-FZF gives a higher 95%-likely SE, which is due to its protective nature of the weak UEs with a lower channel gain. Furthermore, Fig. 6 also shows that the performance gap between L-RZF and L-MMSE is quite small and they outperform the other schemes.

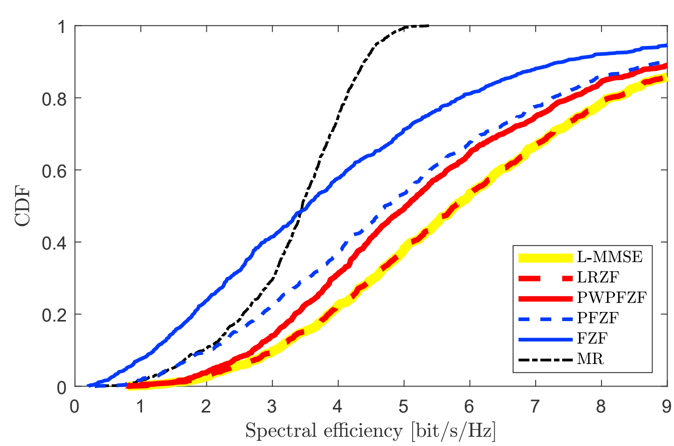


Fig. 6. CDF of the uplink SE per UE achieved by MR [12], L-MMSE [49], and different ZF-like [76] combining schemes.

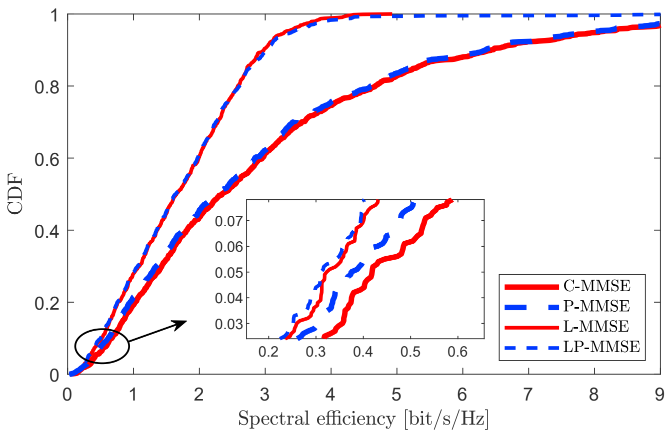


Fig. 7. CDF of the uplink SE per UE achieved by different scalable [21] and non-scalable combining [12] schemes.

Table 5
Computational complexity per AP in terms of number of complex multiplications.

Scheme	Computational Complexity
MR	–
L-RZF, FZF	$\frac{3N\tau_p^2}{2}+\frac{N\tau_p}{2}+\frac{\tau_p^3-\tau_p}{3}$
P-FZF	$\frac{3N\tau_{S_l}^2}{2}+\frac{N\tau_{S_l}}{2}+\frac{\tau_{S_l}^3-\tau_{S_l}}{3}$
PWP-FZF	$\frac{3N\tau_{S_l}^2}{2}+\frac{N\tau_{S_l}}{2}+\frac{\tau_{S_l}^3-\tau_{S_l}}{3}+2(\tau_p-\tau_{S_l})N\tau_{S_l}$
LP-MMSE	$\frac{N^2+N}{2}\lvert\mathcal{D}_l\rvert+\frac{N^3-N}{3}+N^2$
L-MMSE	$\frac{N^2K+NK}{2}+\frac{N^3-N}{3}+N^2$

Fig. 7 shows the CDF of the uplink SE per UE of the scalable centralized P-MMSE and distributed LP-MMSE schemes with two benchmarks where all APs serve all UEs: the C-MMSE combining in (36), and the distributed L-MMSE combining in (37). The first observation is that C-MMSE and P-MMSE outperform L-MMSE and LP-MMSE since the former two schemes exploit much more CSI than the latter two schemes to suppress the interference. Then we can see that the scalable schemes (i.e., P-MMSE and LP-MMSE) provide almost the same performance compared to their counterparts (i.e., C-MMSE and L-MMSE). The negligible performance loss comes from limiting the number of APs serving each UE, which is the price for scalability.

Next, we will first summarize the fronthaul costs, which refer to the amount of information to exchange via the fronthaul network to perform joint coherent transmission/detection and other centralized network operations of different distributed combining schemes. More precisely, when using LSFD, for FZF, P-FZF, PWP-FZF, L-RZF and L-MMSE combining schemes, the number of complex scalars to send from the APs to the CPU via the fronthaul is $(\tau_c-\tau_p)KL$ in each coherence block or $KL+(L^2K^2+KL)/2$ for each realization of the UE locations/statistics.

Then, we summarize the computational complexity with MR, FZF, P-FZF, PWP-FZF, L-RZF, LP-MMSE and L-MMSE combining schemes per AP in terms of the number of complex multiplications in Table 5 according to Refs. [72,76]. Thanks to the fact that $\tau_{\mathcal{T}_l}\le\tau_p$, the complexity of P-FZF and PWP-FZF is lower than that of FZF and L-RZF. Compared with P-FZF, PWP-FZF needs $2(\tau_p-\tau_{\mathcal{T}_l})\tau_{\mathcal{T}_l}N$ more complex multiplications for computing the $\tau_p-\tau_{\mathcal{T}_l}$ MR combining vectors in (49).

3.3. Transmit precoding

In the downlink, the acquired CSI is used to coherently precode the transmitted data signals, which is called *transmit precoding*. Transmit precoding means that each data signal is sent from multiple antennas, but with different amplitudes and phases to direct the signal spatially [53]. Each UE is affected by all the precoding vectors; the own precoding vector is multiplied with the channel response from the serving AP, while the other ones cause interference and are multiplied with the channel response from the corresponding transmitting APs. Hence, the precoding vectors should be selected carefully based on the knowledge of the channel responses [82–84]. In this subsection, we comprehensively introduce the transmit precoding schemes used in CF mMIMO networks.

3.3.1. Precoding via uplink-downlink duality

The precoding vector design is more complicated than that of the combining vector since the downlink SE of UE k depends on the precoding vectors of all UEs in contrast to the uplink SEs that only depend on the UE's own combining vector $\mathbf{v}_k$. The most commonly-used approach for the precoding design is to employ the uplink and downlink duality, which has been introduced in Lemma 4 in Section 2.1. Based on that, all the linear combining schemes mentioned before can be utilized for designing the corresponding precoding vectors [85].

3.3.2. Precoding via utility maximization

The previous method only provides heuristic precoding vectors, and then uses the downlink power allocation to further tune the performance of the UEs. However, an optimal collection of the centralized precoding vectors can be computed by maximizing a system-wide utility function [62,86]. Suppose the utility is to maximize the minimum instantaneous SINR of all UEs. In that case, the optimal solution is obtained by solving a second-order cone program [86], and arbitrary power constraints can be added to that problem [62]. Other metrics, such as sum-rate maximization can be considered but only the local optimum can be solved. The benefit of optimal beamforming is that it outperforms any other schemes and comes with an optimal power allocation, while the drawback is the high computational complexity because each instance of the optimization problem is complex and needs to be solved once per coherence block. This approach cannot be made scalable.

3.3.3. Precoding via over-the-air signaling

The precoding vectors should be optimized jointly for all UEs when not constructed based on the uplink-downlink duality. However, the distributed precoding design is hard to be carried out. The reason is that the optimization of $\mathbf{w}_{kl}$ needs information about the channel conditions between AP k and the other APs and about the precoding vectors adopted by the latter for UE k. Such cross-term information is acquired utilizing the fronthaul and must be adjusted iteratively, which leads to a sizeable fronthaul load. A novel Over-The-Air (OTA) approach which can optimize precoding vectors in a distributed fashion is proposed in Refs. [87,88]. It is based on a particular uplink signaling resource together with a new CSI combining mechanism. In this way, each AP can acquire the cross-term information over the air rather than via extensive fronthaul signaling.

The iterative implementation of the distributed precoding design via OTA signaling is illustrated in Fig. 8. To be specific, as a starting point, each AP initializes its precoding vector. After that, each AP transmits a superposition of the pilots after precoding it with the corresponding precoding vector and each UE receives a downlink pilot signal $\mathbf{y}_{\mathrm{DL}}$ and computes its combining vector $\mathbf{v}_{\mathrm{OTA}}$ based on $\mathbf{y}_{\mathrm{DL}}$. Then, each UE transmits its pilot combined after precoding it with its combining coefficient v_{OTA} and each AP receives $\mathbf{y}_{\text{UL-1}}$. Besides, each UE transmits $v_{\mathrm{OTA}}^{\mathrm{H}}\mathbf{y}_{\mathrm{DL}}$ after precoding it with its combining coefficient v_{OTA} while each AP receives $\mathbf{y}_{\text{UL-2}}$. Finally, each AP computes its precoding vector based on $\mathbf{y}_{\text{UL-1}}$ and $\mathbf{y}_{\text{UL-2}}$. This procedure repeats until a predefined criterion is satisfied. Unlike the conventional precoding design where the CSI among the APs is exchanged via fronthaul signaling, the aforementioned approach exchanges the CSI among the APs via the additional uplink OTA signaling

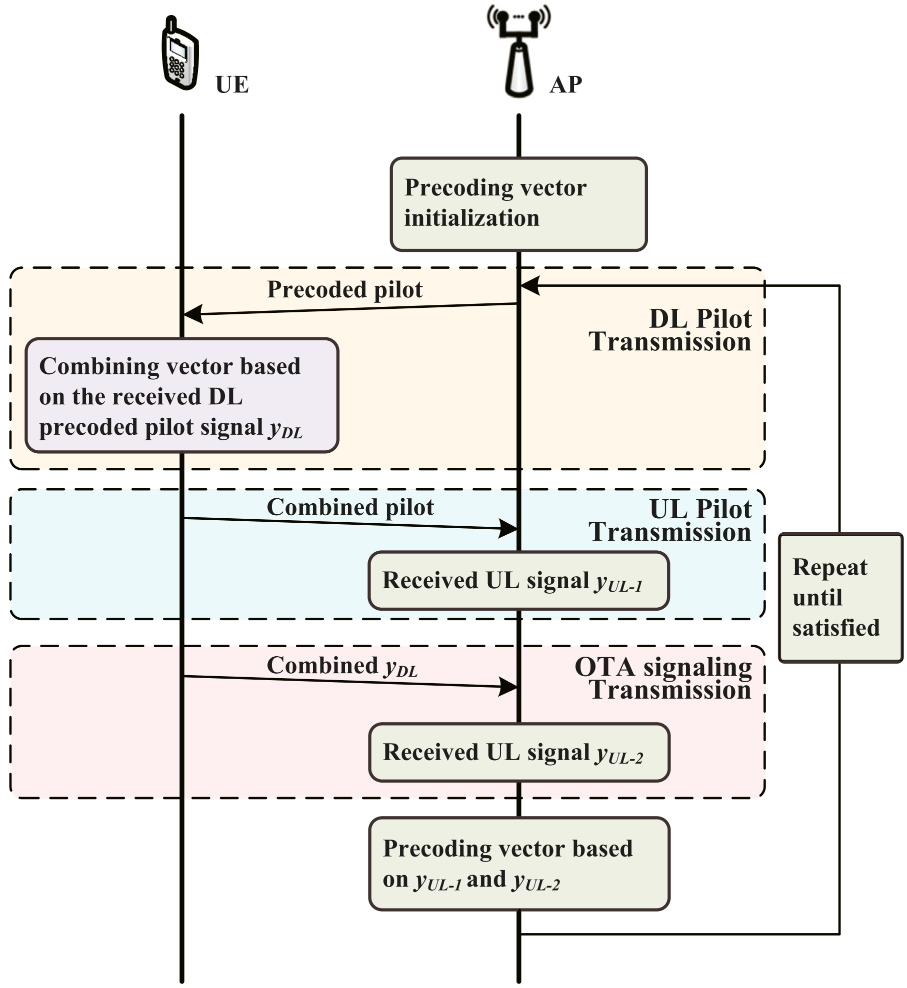


Fig. 8. Flow chart of the OTA-aided precoding design.

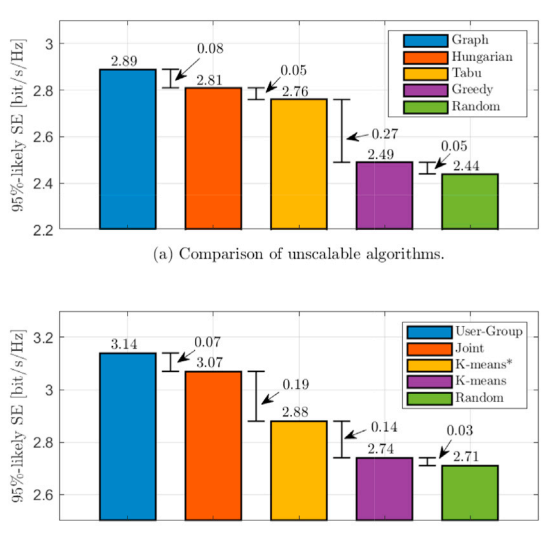


Fig. 9. 95%-likely SE with different pilot assignment algorithms. In abbreviation, the unscalable algorithms: Greedy [51], Tabu [95], Hungarian [96], and Graph [93]. The scalable algorithms: Random [51], K-means [97], K-means* [64], Joint [21], and User-Group [64].

resource (i.e., $v_{\mathrm{OTA}} v_{\mathrm{OTA}}^{\mathrm{H}} y_{\mathrm{DL}}$), which advances in terms of scalability and flexibility (see Fig. 9).

3.4. User access and association

When a UE is about to commence its communication, it first needs to access the network and then be assigned resources for the upcoming signal processing, such as the pilot sequence, the serving APs, etc. In this subsection, we review the emerging schemes for the user access of the CF mMIMO communication network concerning AP selection, pilot assignment, user activity detection, and AP switch on/off strategies.

3.4.1. AP selection

Compared to the cellular networks where each UE is only associated with one AP, the CF systems require more fronthaul connections to transfer each UE's data to/from multiple APs, which leads to extra fronthaul provisioning and energy consumption. However, to avoid substantial pilot contamination, each AP can only serve a limited number of UEs due to the pilot shortage. For the above observations, the original design of CF mMIMO systems in Ref. [51] wherein all UEs in the network are simultaneously served by all APs, is unsuitable for the practical implementation of CF mMIMO. Motivated by this, each UE should not be served by all APs, but by a subset of selected APs, which is the so-called *AP selection* and typically only serve at most UE per pilot.

There are two types of AP selection schemes: large-scale-based scheme [20,89] and competition-based scheme [64]. In the former category, each UE k selects $|\mathcal{M}_k| \leq L$ dominant APs corresponding to the $|\mathcal{M}_k|$ largest large-scale fading coefficients, which satisfy

$$\sum_{l\in\mathcal{M}_k}\frac{\beta_{kl}}{\sum_{j=1}^{L}\beta_{kj}}\geq\delta\% \tag{52}$$

where $\left\{\beta_{kl}\right\}$ is the sorted (in descending order) version of the set $\{\beta_{kl}\}$, and the threshold δ indicates the settled percentage of the total received power that signal APs contribute to each UE. The large-scale-based selection scheme associates UEs and APs in a user-centric manner. However, if an AP serves more than one UE per pilot, the signals from and to these pilot-sharing UEs will be strongly interfering, which is undesired. Hence, an improved AP selection scheme could be designed from a UE perspective, but under constraints set by the APs' capabilities, e.g., an AP can only serve at most τ_p UEs. For that purpose, a competition for an AP l occurs when a new accessing UE k attempts to select AP l while AP l already has τ_p UEs in $\mathcal{D}_l$. The principle of the competition-based selection is that an AP l gives priority to the UEs providing the best channel conditions. Precisely, AP l finds the "weakest" UE

$$k^{*}=\arg\min_{i\in\{k\}\cup\mathcal{D}_l}\beta_{il} \tag{53}$$

If $k^*=k$, UE k puts AP l into its *blacklist*; otherwise, UE k succeeds UE k^* in $\mathcal{D}_l$, and UE k^* puts AP l into its blacklist likewise. If a UE has $L-1$ APs on its blacklist, which means it has lost every competition it participated in. In that case, this UE will be associated with the only AP left and will not participate in another contest. This operation prevents the weak UEs from being abandoned; at the end of the day, pilot-contamination issues can also be dealt with when designing the power allocation, so it is acceptable to occasionally let APs serve more than τ_p UEs and deal with the issue later. The competition-based selection scheme allows the UEs to select as many serving APs as possible to make the best use of the APs service resources and meanwhile prevents an AP from serving more than τ_p UEs. Both of the described methods are heuristic, so there is room for improvements in the future.

3.4.2. Pilot assignment

CSI is essential in multiple antenna systems, both cellular and CF. It is usually acquired through pilot transmission between the UEs and APs. However, the lack of a sufficient number of orthogonal pilot sequences, which comes from the natural channel variations in the time and frequency domains, compels the UEs to reuse the pilot resources, leading to pilot contamination. This phenomenon reduces the channel estimation quality, making coherent transmission less effective and making it harder to reject interference between the pilot-sharing UEs. Thus, a properly designed pilot assignment algorithm is critical to ensure good performance in the CF mMIMO systems.

Random assignment is a well-considered algorithm thanks to its simplicity [46,51], wherein each UE is assigned a fixed pilot at random from the orthogonal pool and uses this pilot during the entire transmission. This simple algorithm is not preferable since the neighboring UEs will occasionally use the same pilot and thus create strong mutual interference that is hard to suppress. A step forward is the greedy algorithm which iteratively updates the pilot of the worst-performing UE after a random assignment [51,90,91]. Greedy algorithms can converge to local optima but are unlikely to provide a globally optimal pilot assignment. Another step is user-centric clustering [92], where the UEs are clustered into groups (joint or disjoint) based on the large-scale information, like the large-scale coefficients, the location of UEs and APs, and the distance between UEs and APs. The pilots are reused in/over these groups. Several algorithms are considered in the clustering pilot assignment [93–96]. Graph coloring is used in Refs. [93,94], where the interference between UEs is modeled as a graph. UEs (shown as vertices in the graph) are connected if at least an AP serves them. Conventional graph coloring algorithms can be exploited to color the UEs with the fewest colors. The final assignment is achieved by updating the interference graph. To avoid being trapped in a local optimum, tabu search is used in Ref. [95], where the *tabu list* records previous assignments to ensure the efficient search of the assignment solution space. The authors of [96] propose an iterative approach based on the Hungarian algorithm. In each iteration, each UE and its neighboring UEs are assigned with mutual orthogonal pilots by exploiting the Hungarian algorithm, given the pilot assignment of the rest of the UEs is fixed. The final assignment is achieved when the performance measures reach convergence, or the iterations reach the allowed maximum number.

Although the aforementioned pilot assignment algorithms limit the pilot contamination to varying degrees, they might not be feasible for practical implementation since the complexity grows polynomially with the number of APs and UEs. Therefore, provably scalable algorithms for pilot assignment have also been developed.

A joint AP selection and pilot assignment algorithm is proposed in Ref. [21], in which a UE first appoints its *Master AP*. Then the Master AP assigns the pilot with the least interference to this UE and informs a limited set of neighboring APs that it is about to serve this UE on the assigned pilot. A neighboring AP decides to serve this UE on the assigned pilot or not based on its serving status. This algorithm achieves scalability by providing each UE with the least bad pilot but performs no optimization for which the pilot assignment is fairly straightforward. The K-means algorithm is used for user clustering in Refs. [64,97], where the UEs are separated into disjoint clusters based on the knowledge of the location and interference relationship of UEs and APs, respectively. The UEs in the same cluster are assigned with mutual orthogonal pilots.

Table 6
Concise description of pilot assignment algorithms.

Scheme	Concise Description	Online Complexity
Random assignment [51]	Each UEs is assigned a fixed pilot from the orthogonal	$\mathcal{O}(K)$
Greedy [51, 90, 91]	Start from a random assignment, iteratively update the pilot of the worst performing UE	$\mathcal{O}(KL)$ [51]
Dynamic Pilot Reuse [92]	Based on user-centric concept, two UEs with a large distance can share the same pilot	$\mathcal{O}(\mathbb{N}\mathbb{T}K^2)$
Graph Coloring [93, 94]	Interference is modeled as a graph and graph coloring algorithm is exploited to assign all UEs with fewest pilots	$\mathcal{O}(K^2+KL+KL\log_2L)$ [93]
Tabu-Search [95]	*Tabu list* prevent the assignment from being trapped in the local optimum and ensure the efficient search of the assignment solution space	$\mathcal{O}(N_{\text{tabu}}K^2L)$
Hungarian Algorithm [96]	Each UE and its neighboring UEs are assigned with mutual orthogonal pilots by exploiting Hungarian Algorithm, given the pilot assignment of the rest UEs is fixed	$\mathcal{O}\left(K\left(L+\tau_p^2\right)\right)$
Joint AP Selection and Pilot Assignment [21]	Each UE pinots its *Master AP*, which assigns the pilot with least interference to this UE and informs the neighboring APs to cooperatively serve this UE with the assigned pilot	$\mathcal{O}(L+K)$
K-Means [64,97]	K-means algorithm is used to separate the UE into disjoint clusters. The UEs in the same cluster are assigned with mutual orthogonal pilots	$\mathcal{O}\left(K^2/\tau_p+\tau_p^2\lceil K/\tau_p-1\rceil\right)$ [64]
User-Group [64]	Interference relationship between the UEs are exploited to iteratively separate the UEs into disjoint groups. The UEs in the same cluster share the same pilot	$\mathcal{O}(K^2L)$

Although the K-means algorithm separates the UE clusters as far as possible, it operates on the cluster level; or in other words, it dynamically divides the network into subareas, but it cannot prevent the neighboring UEs in different subareas from sharing the same pilot. To solve this issue, the authors of [64] proposed a user-group algorithm operating at the UE level, where the interference relationship between the UEs is exploited to separate the UEs into disjoint groups iteratively. The UEs in the same cluster share the same pilot.

A concise description of the pilot assignment algorithms is shown in Table 6, where the online complexity analysis is given. Moreover, a comparison of the pilot assignment algorithms on the 95%-likely SE is illustrated in Fig. 9 with the simulation setup of $L = 100$, $K = 50$, $\tau_c = 200$, $\tau_p = 10$, and area of 0.5×0.5 km^2. Although some algorithms outperform others in this given setup, the situation might change in another simulation setup. Anyway, the algorithms with more complicated processing mechanisms and more information of the APs, UEs, and communication environment conditions will offer better performance. Nevertheless, almost all these above algorithms are either unscalable or heuristic, which encourages us to look into the new scalable alternatives combined with the user-centric clustering approach for the optimized pilot assignment in future works. Machine learning might offer a potential solution thanks to its powerful signal processing ability to deal with the highly loaded networks and the relatively low complexity of online computing and framework.

3.4.3. User activity detection

User activity detection is considered in highly crowded scenarios, like the Internet of Things (IoT) and the Internet of Everything (IoE). One feature of these scenarios is that the devices only send small amounts of data and should be energy efficient. Thus the overhead due to random access and scheduling is extensive compared to the data; the other is the sporadic nature of the transmission, i.e., only a relatively small fraction of UEs stays active with short-length payloads. Both features make the grant-free access scheme a promising solution where the active UEs transmit their pilots and payloads simultaneously without scheduling in advance. As the networks continue to densify, the user activity detection of the grant-free access is about to be a non-negligible problem.

Authors in Ref. [98] formulated user activity detection as a maximum likelihood problem. Based on coordinate descent, a detection algorithm with affordable complexity is provided. Authors in Ref. [99], on the other hand, formulated the user activity detection as a Compressive Sensing (CS) problem by exploiting the angular-domain sparsity of the CF mMIMO channels, where the OFDM technique is used for uplink transmission. Both of the above algorithms employ non-orthogonal pilot sequences for user identification.

Achieving massive user signatures by sacrificing the orthogonality of the pilot sequences will reduce the channel estimation quality, which deteriorates the system SE performance. How to efficiently juggle the massive access and high spectral efficiency is an open issue in the considered highly crowded CF mMIMO systems.

3.4.4. AP switch on/off

With the growing demand for green communications, the AP Switch On/Off (ASO) strategy design is becoming a raising topic in CF mMIMO networks. Some APs are dynamically turned On/Off based on the location and data traffic generated by the served UEs. The rationale behind ASO in CF mMIMO is that a large number of APs are implemented in the network, and their neighboring APs could likely fill the SE requirement of the UEs. The goal of ASO is efficiently exploiting some, not all, competent APs for serving the dynamic traffic load requests to improve the system Energy Efficiency (EE), which will be defined later, and decrease the carbon footprint.

Unlike the other works in CF mMIMO considering a static network where the APs are always active, and the status of them is irrelative with when and where the traffic load requests come from [100–102], treat the status of the APs as an optimization variable to serve the UEs in a more efficiently fashion. A globally optimal solution of ASO is provided in Ref. [100] by solving a mixed-integer Second-Order Cone Program (SOCP). Due to the high computing complexity of the non-convex optimization problem, two heuristic low-complexity algorithms are also proposed by utilizing the channel sparsity structure. Authors in Refs. [101,102] developed a collection of heuristic AP switch On/Off algorithms based on the location and propagation losses between APs and UEs, namely Random Selection ASO (RS-ASO), Chi-square test-based ASO (ChiS-ASO), Kolmogorov-Smirnov test-based ASO (KS-ASO), Logarithmic Statistical Energy ASO (LSE-ASO), Minimum Propagation Losses-aware ASO (MPL-ASO), Optimal EE-based Greedy ASO (OG-ASO), etc. Among them, the algorithms based on goodness-of-fit techniques (i.e., ChiS-ASO, KS-ASO, and LSE-ASO) significantly outperform RS-ASO by trying to match the spatial distribution of active APs to one of the UEs. OG-ASO offers the best SE performance by wielding the knowledge of spatial correlation matrices, power control matrices, and power consumption metrics. However, MPL-ASO makes a good tradeoff between the SE performance and complexity by exploiting the large-scale fading coefficients between the APs and UEs with a minor performance penalty.

3.5. Power control and power allocation

To obtain good system performance, the available radio resources must be efficiently managed. To be specific, the K UEs must select appropriate transmit powers $0 \leq p_k \leq p_{\max}$, $k = 1, \ldots, K$ during the uplink transmission, while the L APs must allocate their transmit power $0 \leq \rho_l \leq \rho_{\max}$, $l = 1, \ldots, L$ during the downlink transmission. The procedure of controlling the uplink transmit power is called *power control*, while the procedure of allocating the downlink transmit power between UEs is called *power allocation*. Ideally, the power control/allocation should be carried out to optimize some system-wide utility function, describing the system performance as a whole. Since the utility function determines the structure of the optimization problem and thereby which approaches can be applied to solve it, we survey the state-of-the-art power control/allocation algorithms for solving the three most common types of utility optimization problems: max-min fairness, max sum SE, and max EE.

3.5.1. Max-min fairness

The importance of the max-min fairness utility was emphasized in the early works [51,52], where the vision of CF mMIMO was to provide uniformly good service over the entire coverage area. The goal is to maximize the lowest SE among all the UEs in the network, which leads to uniform service, while the channel conditions will determine how good that service quality is. Since the SE of UE k is an increasing function of the effective SINR (see (8) and (14)), maximizing the lowest SE is equivalent to maximizing the lowest effective SINR among all the UEs.

There are several instances of the max-min fairness problem that can be shown to be convex or quasi-convex. Thus the optimal solution can be obtained by exploiting the bisection search, and the convex optimization [103], Geometric Programming (GP), or SOCP [51,52,86]. The SOCP formulation dates back to Ref. [104]. The authors in Ref. [103] formulated a weighted max-min power optimization problem in a multigroup multicast CF mMIMO system that could be cast as a quasi-concave problem via a quadratic convex transformation. Moreover, [86] developed an optimum downlink beamforming method in CF mMIMO systems by solving a max-min problem that maximizes the minimum SINR among all UEs.

There are instances of the max-min fairness problem that are non-convex, in which case one can sometimes find a local optimum by *alternating optimization*, which partitions the optimization variables into several sets and cyclically optimizes one at a time while keeping the other variable sets fixed [105,106]. In this way, the original problem can be effectively decomposed into several subproblems, which could be convex and then be cyclically solved by exploiting the approaches applied for the convex cases (e.g., bisection search, GP [75,107–109], or SOCP). For example [107], considered a mixed Quality-of-Service (QoS) problem, where the minimum SE of non-real-time UEs is maximized while the rates of the real-time UEs meet their target rates. The original non-convex problem was decomposed into two sub-problems wherein the GP was exploited to solve the power allocation problem. The same approach was also applied in Refs. [75,108,109] for solving the max-min fairness problem with power constraints. Another effective approach is exploiting the *Successive Convex Approximation (SCA)*, which is an iterative algorithm where the non-convex term is substituted by its convex approximation. A max-min SE problem for power control in IoT scenario was considered in Ref. [110], where the reverse convex functions were iteratively approximated by their first order Taylor approximates, respectively.

Although the aforementioned algorithms optimize the transmit power for all UEs to maximize the lowest SE in a system-wide manner, it is unavoidable that their computational complexities grow unboundedly with the number of the UEs, K, which makes these algorithms unscalable according to the definition in Section 2.4. Hence, distributed and heuristic schemes are needed to obtain the scalable power control in large, practically implementable CF networks. Each device makes a local decision with limited involvement of the other devices.

Fractional power control is a classical heuristic scheme in uplink multiuser systems. The principle of fractional power control is controlling the UE transmit power to compensate for a fraction of the pathloss differences among the UEs that are partially served by the same APs, where UE k selects its uplink to transmit power as [4,63,64,111].

$$p_k = \frac{\min_{i\in\mathcal{S}_k}\left(\sum_{l\in\mathcal{M}_i}\beta_{il}\right)^{\nu}}{\left(\sum_{l\in\mathcal{M}_k}\beta_{kl}\right)^{\nu}} p_{\max} \tag{54}$$

where the exponent $\nu \in [0, 1]$ dictates the power control behavior. The nominator in (54) forces $p_k \in [0, p_{\max}]$. Note that $\sum_{l\in\mathcal{M}_k}\beta_{kl}$ denotes the total channel gain from UE k to the APs that serve it, of which the value is large when UE k is in good channel condition. A larger value of ν encourages each UE to compensate for the variations in the total channel gain among the UEs in $i \in \mathcal{S}_k$, which promotes more fairness. If $\nu = 0$, then all UEs transmit with the maximum power (i.e., $p_k = p_{\max}$), which is the so-called *equal power allocation* or *full power transmission*.

Fraction power allocation can be used in the downlink, in which case AP l selects the downlink power allocation coefficient for UE k proportionally to the channel gain, β_{kl}, as [4,21].

$$\rho_{kl} = \begin{cases} \dfrac{(\beta_{kl})^{\upsilon}}{\sum\limits_{i\in\mathcal{D}_l}(\beta_{il})^{\upsilon}}\rho_{\max} & \text{if } k \in \mathcal{D}_l \\ 0 & \text{otherwise} \end{cases} \tag{55}$$

where the exponent $\upsilon \in [0, 1]$ dictates the power control behavior. A larger value of υ gives a higher emphasis to the UEs according to their respective channel gains while $\upsilon = 0$ indicates that each UE in $\mathcal{D}_l$ is allocated with equal power $\rho_{kl} = \frac{\rho_{\max}}{|\mathcal{D}_l|}$. This leads to allocating more power to the UEs in better channel conditions, which seems to be contrary to the max-min fairness concept. Still, it is generally not since the UEs in good channel conditions will then avoid high interference. Since the essence of fraction power control/allocation is to impose a structure with a parameter (ν or υ) that can be tuned, we can also apply it to other utility functions and adjust the parameter to identify a suitable heuristic solution.

Apart from traditional optimization and heuristic methods, ML can be utilized to design power control/allocation methods. Such an approach cannot provide a better solution than the one found by classical optimization methods, but it could potentially lower the computational performance [66]. For instance, one can greatly reduce the online computational complexity at the price of offline training [112]. There are ML-based schemes proposed to solve the max-min fairness problem in CF mMIMO [112–114]. Moreover [113], proposed to approximately solve the max-min fairness problem using the local CSI by training a neural network to identify a mapping between that local CSI and the optimal solution to the system-wide max-min fairness problem. An unsupervised learning approach was studied in Ref. [114], which took the large-scale fading coefficients as inputs to learning the map between these coefficients and soft max-min and max-prod power control policies.

3.5.2. Max sum SE

A potential drawback of the max-min SE fairness problem is that a few UEs might drag down the overall system performance with bad channel conditions. The overwhelming majority of UEs in a large network can likely achieve substantially larger SEs while barely affecting the UEs in the worst conditions since every UE only causes interference to a small subset of neighboring UEs. This motivates the maximization of the sum SE, which represents the overall SE performance of the network instead of the SE achieved by a specific UE.

The max sum SE problem is usually not convex (see (8) and (14)). Hence, it is hard to obtain the optimal solution, and we typically need to settle for a local optimum. The aforementioned alternating optimization can be applied for addressing the non-convexity. The classical weighted MMSE method can be utilized to optimize the data power [4], when the pilot power is fixed. The problem of joint data and pilot transmit power control was considered in Ref. [115] using the Lagrange multiplier method. The SCA method can also be applied to maximize sum SE by employing convex optimization where the non-convex term is substituted by its convex approximation [116–120]. A max sum SE optimization problem for a downlink setup with hardware impairment was considered in Ref. [118]. The problem was non-convex and, thus, the SCA policy was employed by reformulating the original problem as an SOCP. The same approach was also applied in Refs. [119,120] when short-term power constraints and low-resolution Analog-to-Digital Converters (ADCs) were considered, respectively.

Since the aforementioned methods are searching for a local optimum, the ML-based schemes can potentially find better solutions [121–123]. A max sum SE problem in an uplink CF mMIMO system was studied in Ref. [121] using Artificial Neural Networks (ANNs), in which the UE positions were taken as input and the power control policy as output. In Ref. [122], a Deep Convolutional Neural Network (DCNN) was considered in an uplink CF mMIMO system with limited-fronthaul, where the LSF information was exploited to predict the max sum SE power control policy. A deep neural network-based power control method was proposed in Ref. [123].

3.5.3. Max EE

When designing a large CF mMIMO network, the EE is another essential performance metric to consider, which indicates not only "how much and fast" the information can be transmitted, but also "how economically" in terms of the energy, i.e., how much energy it takes to reliably transmit a certain amount of information [125,128,129]. Technically, the EE is defined as [12].

$$\mathsf{EE} = \frac{B\cdot\sum_{k=1}^{K}\mathsf{SE}_k}{P_{\text{total}}} \tag{56}$$

where B is the system bandwidth, and P_{total} is the total power consumption which usually includes four main terms: the transmit power, a term accounting for the analog processing the transceiver chains, a term accounting for the digital signal processing, and a term for the fronthaul connections.

The max EE problem in (56) is non-convex, thus the alternating optimization [124] and SCA [125,126] methods are normally used to solve the problem. The EE maximization problem in an mmWave CF mMIMO system was considered in Ref. [124]. Since it is a non-convex problem, the successive power-bound maximization method, which is based on the idea of merging the alternating optimization and sequential convex programming, is used for alternatively optimizing the transmit power of each AP while keeping the transmit power of other APs fixed. Moreover, [126] studied the EE maximization problem with quantization, considering the per-UE power, fronthaul capacity, and throughput requirement constraints. To solve the non-convex problem, the original SCA problem was decoupled into two sub-problems, namely, receiver filter coefficient design and power control. The former was formulated as a generalized eigenvalue problem, while the GP addressed the latter after exploiting an SCA and a heuristic sub-optimal scheme. The SCA method was also exploited in Ref. [125] to maximize the total EE under the per-AP and per-UE power constraints. The problem was approximately solved via a sequence of SOCP. Although the second-order optimization methods have performed very well, their complexities do not scale favorably with the network size, which motivates the first-order methods. Finally, [127] proposed a first-order method for the non-convex programming to the EE maximization problem. This first-order method could achieve the same performance with a faster run time than the second-order methods.

In Table 7, we summarize the optimization utilities, power optimization methods, and the main literature.

4. Practical implementation

Although many sophisticated schemes and algorithms have been designed for CF mMIMO systems, they are mainly based on simplifying assumptions such as error-free fronthaul connections and perfect hardware, which are unlikely to hold in practical deployment. Hence, in this section, we will discuss the practical issues that a CF mMIMO system has to face to implement (i.e., fronthaul limitation and hardware impairment) and survey the existing schemes to address them. Although synchronization is also an essential issue for practical implementation in CF mMIMO systems, little work has been done on this topic. Hence, we will discuss this part in the Future Research Directions section.

4.1. Fronthaul

One of the main issues for CF mMIMO systems is the limited capacity of the fronthaul links from the APs to the CPU [130–132]. Due to a large number of antennas at the APs, a large number of signals should be exchanged between APs and the CPU through the fronthaul links and hence cause huge power consumption (e.g., 0.25 W/(Gbits/s) when using the optical fiber cables [128]). Besides, when converted to the digital form, it requires a huge capacity for the fronthaul links many times the corresponding user data rate in the uplink to ensure signals are transferred with sufficient precision [133]. In the C-RAN literature, this has been estimated as 20–50 times the corresponding data rate, implemented using the Common Public Radio Interface (CPRI) standard [134], typically over optical fiber. Therefore, reducing the fronthaul load constitutes one of the most substantial challenges in practical CF mMIMO systems [130,135–137].

There are two approaches to address the capacity-limited fronthaul issue: 1) quantizing the transmit signals; 2) using structured lattice codes.

Table 7
Power control/allocation approaches.

Utility Function	Max-min Fairness	Max sum SE	Max EE
Alternative Optimization	[75, 105–109]	[115]	[124]
SCA	[110]	[116–120]	[125, 126]
Bisection	[103]	–	–
GP	[75, 107–109]	–	[126]
SOCP	[86]	[118–120]	[125]
Fractional	[21, 63, 64, 111]		
Others	–	[115]	[127]
ML-based	[112–114]	[121–123]	–

4.1.1. Quantizing

Using a small number of bits to quantize the transmit signals is feasible to reduce the fronthaul load. Therefore, employing low-resolution ADCs at APs is a promising and practical solution, facilitating low power consumption and small hardware cost. According to the way the APs process and forward the signals to the CPU, there are four main types of transmission in the uplink:

Compress-Forward-Estimate (CFE) [133,137–142] Each AP compresses the received pilot and data signals separately and forwards the compressed versions over the fronthaul link to the CPU. Then, the channel estimation, the design of combining vectors, and the data recovery are carried out at the CPU. This is a way to implement centralized combining methods over a limited fronthaul network by compression at the APs.

Estimate-Compress-Forward (ECF) [138,139] First, the channel estimation is performed at each AP. Each AP separately compresses the estimated channels and data signals and forwards the former to the CPU. Finally, the CPU recovers the CSI and performs data detection using the centralized combining. Since the compression is implemented at the APs, ECF reduces the fronthaul load in a distributed fashion.

Estimate-Multiply-Compress-Forward (EMCF) [143] Each AP first estimates the channels, multiplies the received data signal by the local combining vector (computed based on the local channel estimate) and compresses and forwards the results to the CPU. Thus, the CPU only performs data detection. Therefore, when using EMCF, the design of combining vectors and the compression is implemented at the APs in a distributed way.

Estimate-Multiply-Compress-Forward-Weight (EMCFW) [141–143] Similar to EMCF, the signal is further multiplied by the receiver filter coefficients at the CPU to improve the performance. The design of the combining vectors and the compression is also implemented in a distributed way.

When it comes to the downlink, two approaches have been considered:

Compress-After-Precoding (CAP) [144–146] The centralized signal with precoding is first computed and then compressed at the CPU before being sent to the APs, which makes CAP suitable for the centralized precoding.

Precoding-After-Compress (PAC) [147] A simple compression is done at the CPU where the symbol for each of the UEs is separately quantized. Then, each AP receives the symbols and designs the precoding vectors for each UE, which makes PAC suitable for the distributed precoding.

4.1.2. Compute-and-forward

Another possible solution could be the Compute-and-Forward (C&F) schemes [148,149], which can reduce the fronthaul load efficiently by decreasing the cardinality of symbols transmitted to the CPU. As shown in Fig. 10, in the C&F scheme, each AP forwards an integer-linear

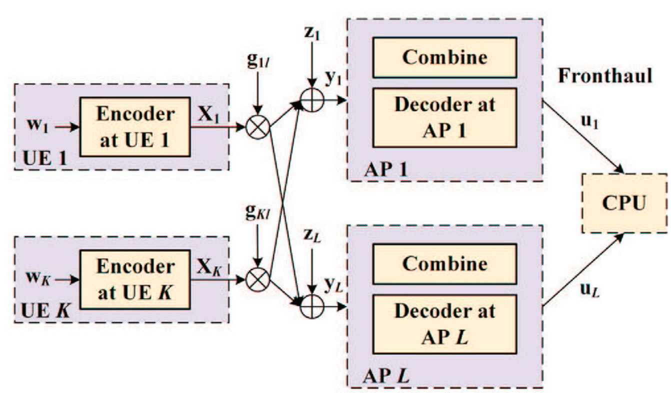


Fig. 10. A C&F framework for CF mMIMO systems.

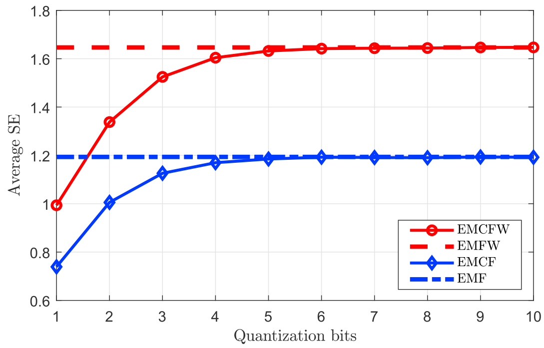


Fig. 11. CDF of the achievable sum-rate with C&F [149] and E-C&F [150] schemes.

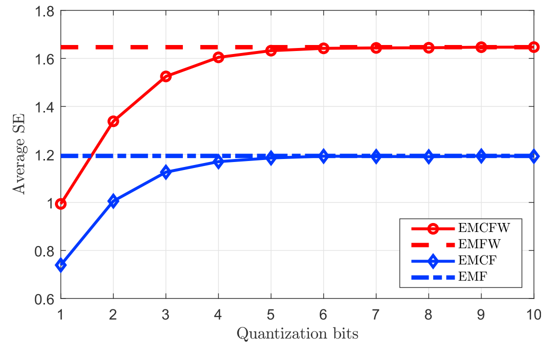


Fig. 12. Average SE per UE versus the number of quantization bits with EMCF [143], EMF, EMCFW [141], and EMFW schemes.

combination of the transmitted signals of all UEs with the same cardinality as each UE's signal. The served UEs can be determined by selecting the coefficient vector and are limited by the computation rate. Although the C&F scheme can obtain the theoretical minimum fronthaul requirement of CF mMIMO with achieving lossless transmission, it is only suitable for the symmetric scenario where all UEs transmit with equal power. However, in CF mMIMO, the UEs can be allocated with unequal transmit power to compensate for pathloss differences. Hence, the Expanded Compute-and-Forward (E-C&F) scheme [150], which is versatile to distribute UE's power unequally and can tolerate different noises at the targeted AP for ensuring the minimum loss in the computation rate, offers significant performance improvement over the C&F scheme. Fig. 11 shows the CDF of the achievable sum-rate obtained via C&F and E-C&F schemes versus the number of APs with $K = 8$, $L = 20$, and $p = 200$ mW. The E-C&F scheme outperforms the conventional C&F scheme in terms of an achievable sum-rate noticeable since the E-C&F framework enables the optimal transmit power of UEs, which facilitates the exploitation of performance gain.

Although quantizing is an efficient and straightforward solution to reducing the fronthaul load, the uniform quantization method can induce the quantization error and achieves a considerable performance loss. Besides, it is shown that performance of the different types of transmission are not the same because they require different fronthaul rate allocations for CSI and/or data signals transmitted to the CPU, and their AP signal processing capabilities are different.

To compare the six transmission strategies by highlighting the required processing at APs and the CPU, Table 8 is provided. Besides, in Fig. 12, we plot the average SE with MR combining at the APs, as a function of the number of the quantization bits for EMCF, EMF, EMCFW, and EMFW, where EMF and EMFW refer to the case of using perfect ADCs in EMCF and EMCFW, respectively, and hence no compression is done at the APs. It can be seen from Fig. 12 that EMCFW provides a larger SE than EMCF. This can be explained that in the optimized receiver filter coefficients step in EMCFW maximizes the SNR. Besides, via numerical results in Refs. [141,143], it is shown that the EMCF can outperform the other schemes when they apply UatF bounding (not the CSI-based ones), and ECF strategy also outperforms the CFE. Thus, to decrease the deployment cost of the network, dummy APs which only compress and forward the received signals are preferred at the cost of performance loss. So, there is a tradeoff between the performance gain and implementation costs which must be taken into account.

4.2. Hardware impairments

To enable a ubiquitous deployment of CF mMIMO, a tradeoff between the cost and the quality of the transceiver hardware in CF mMIMO should be considered since many antenna elements in APs are deployed, which might increase the deployment cost and energy consumption. A possible countermeasure is to make use of the compact low-cost components, which introduce power amplifier non-linearities, phase noise in local oscillators, amplitude/phase imbalance in I/Q mixers, and finite-resolution quantization in ADCs. All these non-idealities are referred to as *Hardware Impairments* (HIs). Most existing works on CF mMIMO neglect the impacts of the HIs by modeling the wireless communication channels as linear filters, as shown in Fig. 13(a). Although this ideal model can be used to devise analog or digital compensation algorithms that could substantially mitigate the impacts of HIs, the residual HIs will still exist due to modeling inaccuracies and the destructive nature of some HIs. Instead, the non-ideal transceiver hardware can be modeled as non-linear memoryless filters to provide better insights for the practical implementation, as shown in Fig. 13(b).

Research on the impact that the residual HIs have on the data rate performance of the CF mMIMO systems has been made in Refs. [129, 151–156]. To be specific, authors in Ref. [129] quantified the performance of both the uplink and downlink CF mMIMO with the classical additive hardware distortion model. By exploiting the MR processing, SE

Table 8
Comparison of the quantization schemes.

Scheme	Processing at the APs	Processing at the CPU	Comb./prec. Design	Compression
CFE [133,137–142]	CSI compression Data compression	Channel estimation Combining design	Centralized	Distributed
ECF [138,139]	Channel estimation CSI compression Data compression	Combining design	Centralized	Distributed
EMCF [143]	Channel estimation Combining design	–	Distributed	Distributed
EMCFW [141–143]	Channel estimation Combining design	Receiver filter design	Distributed	Distributed
CAP [144–146]	Precoding design Compression	–	Centralized	Centralized
PAC [147]	Compression	Precoding design	Distributed	Distributed

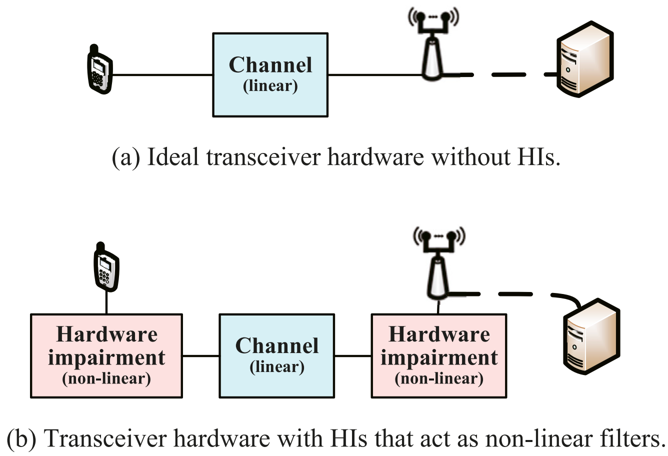


Fig. 13. A generalized CF mMIMO system with/without HIs.

and EE were derived in closed-form. With these tractable expressions, hardware-quality scaling laws were presented, which proved that the detrimental effect of the HIs at the APs vanishes as the number of APs increases. Moreover, a max-min power control algorithm was proposed to maximize the minimum UE data rate. Mitigating the impacts of HIs by using different levels of AP cooperation was investigated in Ref. [151], where the APs could perform data decoding fully distributively or by exploiting LSFD. The results revealed that the LSFD could provide the largest SE under the HIs. Authors in Ref. [152] considered the HI in analyzing the uplink SE for the wireless powered UAV communication in CF mMIMO systems [153,154]. focused on the impact of employing low-resolution ADCs on the CF mMIMO. A simple asymptotic approximation for the achievable rate was derived by considering the effects of AP and UE number, antenna number per AP, and ADC resolution. The results showed that the achievable rate of an arbitrary UE converges to a finite limit, independent of the ADC resolution of the APs, as the number of APs goes to infinity. Additionally, an ADC resolution bits allocation scheme was proposed to maximize the sum rate given a fixed total ADC resolution bits. Authors in Ref. [155] considered the physical layer security in a CF mMIMO system with HIs, where a lower bound for the ergodic secrecy rate in the presence of pilot spoofing attack and imperfect CSI is derived. Moreover, an optimal power allocation scheme was obtained to maximize the achievable secrecy rate using the continuous approximation and path-following algorithms. Analytical results revealed that the hardware-quality scaling law is almost inapplicable for secure transmission in the CF mMIMO system except for some particular scaling factors. An article that predates the CF mMIMO area but applies to the same scenario is [156], which assessed the impact of HIs on scalable CF mMIMO systems by also considering the effect of phase noise.

5. Future research directions

In this section, we briefly highlight some major open problems and research challenges to be addressed in future work.

5.1. Multiple CPUs and practical fronthaul topology

Much of the algorithmic design for CF mMIMO has been developed to be transparent to the topology of the underlying network architecture [4], to make it applicable when having one or multiple CPUs, and having parallel or sequential fronthaul connections. In the canonical case, the network comprises many distributed APs with independent cables to the single CPU, which is also known as a star topology. Still, it is unlikely that geographically large networks will be deployed in that manner. There might be multiple CPUs that are connected to disjoint subsets of the APs [157], thus if two APs that belong to different CPUs are cooperating, the fronthaul signaling will have to involve multiple APs. Moreover, the radio-stripes topology has been proposed in Ref. [46], where a set of APs are deployed along a fronthaul cable and, thus, have a sequential connection to the CPU. This design is motivated by the practical need for limiting the total cable lengths and opens up research questions related to how the communication algorithms can be adapted to exploit the finer details of the fronthaul topology. For example, some centralized processing schemes can be implemented sequentially [158]. There are many open research challenges related to distributing the signal processing over multiple CPUs and adapting the algorithms to exploit the structure.

Each fronthaul connection will have a limited capacity. The leading theory for CF mMIMO has been developed under the assumption of infinite fronthaul connections, but with a general awareness that one needs to limit the number of signals transmitted between APs and CPUs to achieve scalability [4]. When dealing with a practical capacity-limited fronthaul, one must consider the tradeoff between precision and the number of conveying signals. As discussed in Section 4.1, it then matters where the processing is done: Signals that are measured at the APs can either be processed there at full precision or elsewhere with reduced precision. When it comes to signal compression for fronthaul signaling, one can either consider model-aided or data-driven methods, where the latter can make use of autoencoder methodologies.

5.2. Synchronization

Coherent signal processing is possible only if the APs maintain a sufficiently accurate relative timing and phase synchronization. The network might have an absolute time (phase) reference, but the APs are unsynchronized. Therefore, AP synchronization and TDD reciprocity calibration are two critical problems to enable CF mMIMO. Suppose each AP has a local oscillator, and the wired fronthaul network cannot provide a sufficiently accurate common time and frequency reference. In that case, such synchronization must occur via OTA signaling [46]. AirSync, which provides timing and phase synchronization accuracy, has been implemented in the distributed mMIMO [159]. Specifically, it detects the slot boundary such that all APs are time-synchronous within a Cyclic Prefix (CP) of the OFDM modulation and predicts the instantaneous carrier phase correction along with the transmit slot such that all transmitters maintain their coherence. To limit the reciprocity and synchronization errors, a synchronization process needs to be applied at regular intervals. High-precision inter-node clock synchronization is a prerequisite for joint processing of the distributed mMIMO [160]. All Radio Access Unit (RAU) clocks in the system are assigned by the master node through IEEE 1588 PTPv2. In practice, a Global Position System (GPS) can also be used for more precise synchronization. More generally, the communication theory that underpins CF mMIMO assumes a perfect timing synchronization, which is physically impossible over a large network, even if the clocks are synchronized. Hence, there is room for theoretical advancements as well.

5.3. Mobile edge computing

The previous subsections focused on where they carry out the lower-layer processing in a communication network. A related concept is mobile edge computing (MEC), where the computation/storage resources of the higher layers in the network are pushed to the edge to alleviate the burden of core networks [161,162]. MEC will naturally reduce the latency since processing is moved closer to the UEs and use general-purpose cloud computing hardware that can be co-located with CPUs. The user-edge-cloud architecture conceived for MEC perfectly matches the UE-AP-CPU architecture of CF mMIMO, making MEC and CF mMIMO a perfect fit [99,163]. Edge nodes (APs and CPUs) equipped with computation/storage capability could deal with UEs' computation and content requests, and consequently, reduce the transmission delay and requirement of the fronthaul/backhaul connection capacities. The research into this direction is in its infancy.

5.4. Enabling federated learning

Apart from using the downlink to improve the channel estimation or resource allocation in the CF mMIMO system, as described earlier in this survey, a wireless network can also be part of the infrastructure used when implementing machine learning algorithms for other applications. The Federated Learning (FL) concept can facilitate the collaborative learning processes of complex models among the distributed devices, keeping their local training data and control privacy. The signals sent from the UEs over the wireless network are suggested local model updates (i.e., the model's weights), aggregated at the core of the network, where global model updates are determined and broadcasted to the UEs. The distributed processing manner of FL naturally fits CF mMIMO, which makes CF mMIMO an enabler of FL [164,165].

5.5. Multi-antenna UEs

The main theory for CF mMIMO has been developed for single-antenna UEs, even though contemporary UEs have at least two antennas, and future devices will feature even larger arrays when operating in the mmWave bands. First steps towards considering CF mMIMO with multi-antenna UEs are found in Refs. [73,83,166]. In general, the multiple antennas can either be used for spatial multiplexing of multiple streams per UE (up to one per antenna) or for improved precoding/combining that mitigates interference [167]. While the achievable SE can be quantified using existing methods, there are many open resource allocation questions related to power allocation, pilot assignment, and precoding/combining design.

5.6. Channel estimation and prediction beyond the block fading model

In practice, the wireless channels vary continuously over time and frequency, not in the block-fading manner described in this survey and most of the theoretical works on this topic. On the one hand, the coherence block size can always be dimensioned in a conservative manner such that the channels are indeed approximately constant within each block. On the other hand, underlying physical rules dictate how the channel can evolve frequently. By exploiting such properties, using model-aided or data-driven approaches [168], the communication performance can be significantly improved: pilots can be transmitted less frequently, and/or the CSI quality can be increased. Prior work on this topic has been done in the cellular mMIMO field [169], which potentially can be adapted to cover CF mMIMO.

5.7. Integrated sensing and computing

In future wireless communications, Integrated Sensing And Communication (ISAC) will be a paradigm change. Some promising ISAC-like Dual-Functional Radar-Communication (DFRC) system has attracted substantial attention, where joint radar sensing and multi-user communication can be simultaneously implemented [170,171]. And it is also interesting to investigate a DFRC-based system with CF mMIMO for the feature of uniform coverage. However, signal processing and system design will be the key challenges. Besides, the great demand for computation is also an issue to address. Therefore, the DFRC system based on scalable CF mMIMO will be investigated in the future.

6. Conclusion and lessons learned

Exploiting densification and decentralization to boost the user-experienced data rates and realize a ubiquitous service is an irresistible general trend for future wireless communications. CF mMIMO represents an attempt to reach this promising prospect by coordinating dense serving antennas in a decentralized CF approach, which greatly squeezes the potential of the multiple antenna technology so that it dynamically achieves the best performance with the available resources. In particular, the dense deployment of the serving antennas in CF mMIMO will result in strong macro-diversity from a significantly smaller average distance between a UE and its closest antennas, while the joint signal processing and scheduling among the distributed antennas achieve array gains and spatial interference suppression which substantially reduce the QoS variations within the coverage area.

In this paper, we present a comprehensive review of the concepts and techniques proposed for CF mMIMO systems. First, we give the motivation for CF mMIMO and provide a brief introduction of CF mMIMO itself and the other technologies related to it. Then we use a section to briefly provide a tutorial about the technical foundations of CF mMIMO, including the transmission procedure and mathematical system model. The core of the paper provides an extensive survey on the state-of-the-art schemes and algorithms available in the literature for resource allocation and signal processing (i.e., channel estimation, combining and precoding, user access and association, and power control) and practical implementation (i.e., fronthaul limitation and hardware impairment) in CF mMIMO systems. We then highlight the open research areas for CF mMIMO (e.g., multiple CPUs cooperation, MEC, enabling machining learning, etc.) and propose some potential approaches for solutions.

Although the research on this topic is still in the exploratory phase, the primary demonstrations, field tests, and prototypes of CF mMIMO systems have been ongoing across different projects in academia and industry. How to realize the scalable intelligent system deployment with new mathematical tools, new applications, and new standardizations becomes a very attractive open issue for all researchers in this field. Though many challenges remain to address, CF mMIMO shows great potential to meet the ubiquitous high QoS demands of the 6G communications. In the foreseeable future, research on CF mMIMO will continue to mature. With no doubt, this technology with its concepts will open up new frontiers in wireless services and applications.

Declaration of competing interest

We have no conflicts of interest to disclose.

Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant 2020YFB1807201, in part by National Natural Science Foundation of China under Grants 61 971 027, U1834210, and 61 961 130 391, in part by Beijing Natural Science Foundation under Grant L202013, in part by Frontiers Science Center for Smart High-speed Railway System under Grant 2020JBZD005, in part by the Royal Society Newton Advanced Fellowship under Grant NA191006. E. Björnson was supported by the Grant 2019–05 068 from the Swedish Research Council.

References

[1] ITU, Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s), Tech. rep., ITU-R M.2410-0, 2017.
[2] M. Cooper, The Myth of Spectrum Scarcity, Tech. rep., DYNA llc, Mar. 2010 https://ecfsapi.fcc.gov/file/7020396128.pdf (Accessed 10 March 2021).
[3] V.H. MacDonald, The cellular concept, Bell Syst. Tech. J. 58 (1) (1979) 15–41.
[4] Ö.T. Demir, E. Björnson, L. Sanguinetti, Foundations of user-centric cell-free massive MIMO, Found. Trends® Signal Process. 14 (3–4) (2021) 162–472.
[5] J. Hoydis, M. Kobayashi, M. Debbah, Green small-cell networks, IEEE Veh. Technol. Mag. 6 (1) (2011) 37–43.
[6] I. Hwang, B. Song, S.S. Soliman, A holistic view on hyper-dense heterogeneous and small cell networks, IEEE Commun. Mag. 51 (6) (2013) 20–27.
[7] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, T. Svensson, The role of small cells, coordinated multipoint, and massive mimo in 5G, IEEE Commun. Mag. 52 (5) (2014) 44–51.
[8] D. López-Pérez, M. Ding, H. Claussen, A.H. Jafari, Towards 1 Gbps/UE in cellular systems: understanding ultra-dense small cell deployments, IEEE Commun. Surv. Tutor. 17 (4) (2015) 2078–2101.
[9] J.G. Andrews, S. Buzzi, W. Choi, S.V. Hanly, A. Lozano, A.C. Soong, J.C. Zhang, What will 5G be? IEEE J. Sel. Area. Commun. 32 (6) (2014) 1065–1082.
[10] S. Parkvall, E. Dahlman, A. Furuskär, M. Frenne, NR: the new 5G radio access technology, IEEE Commun. Stand. Mag. 1 (4) (2017) 24–30.

[11] T.L. Marzetta, Noncooperative cellular wireless with unlimited numbers of base station antennas, IEEE Trans. Wireless Commun. 9 (11) (2010) 3590–3600.
[12] E. Björnson, J. Hoydis, L. Sanguinetti, Massive MIMO networks: spectral, energy, and hardware efficiency, Found. Trends® Signal Process. 11 (3–4) (2017) 154–655.
[13] J. Zhang, E. Björnson, M. Matthaiou, D.W.K. Ng, H. Yang, D.J. Love, Prospective multiple antenna technologies for beyond 5G, IEEE J. Sel. Area. Commun. 38 (8) (2020) 1637–1660.
[14] J.G. Andrews, X. Zhang, G.D. Durgin, A.K. Gupta, Are we approaching the fundamental limits of wireless network densification? IEEE Commun. Mag. 54 (10) (2016) 184–190.
[15] S. Chen, F. Qin, B. Hu, X. Li, Z. Chen, User-centric ultra-dense networks for 5G: challenges, methodologies, and directions, IEEE Wireless Commun 23 (2) (2016) 78–85.
[16] M. Kamel, W. Hamouda, A. Youssef, Ultra-dense networks: a survey, IEEE Commun. Surv. Tutor. 18 (4) (2016) 2522–2545.
[17] E. Björnson, J. Hoydis, L. Sanguinetti, Massive MIMO has unlimited capacity, IEEE Trans. Wireless Commun. 17 (1) (2017) 574–590.
[18] H.Q. Ngo, A. Ashikhmin, H. Yang, E.G. Larsson, T.L. Marzetta, Cell-free massive MIMO: uniformly great service for everyone, in: Proc, IEEE SPAWC, 2015, pp. 201–205.
[19] S. Buzzi, C. D'Andrea, User-centric communications versus cell-free massive MIMO for 5G cellular networks, in: WSA 2017; 21th International ITG Workshop on Smart Antennas, VDE, 2017, pp. 1–6.
[20] S. Buzzi, C. D'Andrea, Cell-free massive MIMO: user-centric approach, IEEE Wireless Commun. Lett. 6 (6) (2017) 706–709.
[21] E. Björnson, L. Sanguinetti, Scalable cell-free massive MIMO systems, IEEE Trans. Commun. 68 (7) (2020) 4247–4261.
[22] C.M.R. Institute, C-RAN the Road towards Green RAN, White paper, Oct. 2011.
[23] A.D. Wyner, Shannon-theoretic approach to a Gaussian cellular multiple-access channel, IEEE Trans. Inf. Theor. 40 (6) (1994) 1713–1727.
[24] S. Shamai, B.M. Zaidel, Enhancing the cellular downlink capacity via co-processing at the transmitting end, in: IEEE VTC-Spring, 3, 2001, pp. 1745–1749.
[25] S. Zhou, M. Zhao, X. Xu, J. Wang, Y. Yao, Distributed wireless communication system: a new architecture for future public wireless access, IEEE Commun. Mag. 41 (3) (2003) 108–113.
[26] S. Venkatesan, A. Lozano, R. Valenzuela, M.I.M.O. Network, Overcoming intercell interference in indoor wireless systems, in: Asilomar Conf. Signals Syst. Comput., 2007, pp. 83–87.
[27] S.A. Jafar, G.J. Foschini, A.J. Goldsmith, Phantomnet: exploring optimal multicellular multiple antenna systems, EURASIP J. Wirel. Commun. Netw. 2004 (5) (2004) 1–14.
[28] H. Zhang, H. Dai, Cochannel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks, EURASIP J. Wirel. Commun. Netw. 2 (2004) 222–235.
[29] P. Zhang, X. Tao, J. Zhang, Y. Wang, L. Li, Y. Wang, A vision from the future: beyond 3G TDD, IEEE Commun. Mag. 43 (1) (2005) 38–44.
[30] X. Tao, J. Xu, X. Xu, C. Tang, P. Zhang, Group cell FuTURE B3G TDD system, in: IEEE Ann. Int. Symp. Pers., Indoor, and Mobile Radio Commun. (PIMRC) 2, 2005, pp. 967–971.
[31] G. Foschini, K. Karakayali, R. Valenzuela, Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency, IEE Proc. Commun. 153 (4) (2006) 548–555.
[32] M.K. Karakayali, G.J. Foschini, R.A. Valenzuela, Network coordination for spectrally efficient communications in cellular systems, IEEE Wireless Commun 13 (4) (2006) 56–61.
[33] A. Tolli, M. Codreanu, M. Juntti, Cooperative MIMO-OFDM cellular system with soft handover between distributed base station antennas, IEEE Trans. Wireless Commun. 7 (4) (2008) 1428–1440.
[34] A. Sanderovich, O. Somekh, H.V. Poor, S. Shamai, Uplink macro diversity of limited backhaul cellular network, IEEE Trans. Inf. Theor. 55 (8) (2009) 3457–3478.
[35] O. Simeone, O. Somekh, H.V. Poor, S. Shamai, Downlink multicell processing with limited-backhaul capacity, EURASIP J. Adv. Signal Process. (2009) 1–10, 2009.
[36] S. Khattak, W. Rave, G. Fettweis, Distributed iterative multiuser detection through base station cooperation, EURASIP J. Wirel. Commun. Netw. (2008) 1–15, 2008.
[37] E. Björnson, R. Zakhour, D. Gesbert, B. Ottersten, Cooperative multicell precoding: rate region characterization and distributed strategies with instantaneous and statistical CSI, IEEE Trans. Signal Process. 58 (8) (2010) 4298–4310.
[38] D. Gesbert, S. Hanly, H. Huang, S.S. Shitz, O. Simeone, W. Yu, Multi-cell mimo cooperative networks: a new look at interference, IEEE J. Sel. Area. Commun. 28 (9) (2010) 1380–1408.
[39] S. Parkvall, E. Dahlman, A. Furuskar, Y. Jading, M. Olsson, S. Wanstedt, K. Zangi, LTE-advanced - evolving LTE towards IMT-advanced, in: IEEE VTC-Fall, IEEE, 2008, pp. 1–5.
[40] M. Boldi, A. Tölli, M. Olsson, E. Hardouin, T. Svensson, F. Boccardi, L. Thiele, V. Jungnickel, Coordinated multipoint (CoMP) systems, in: A. Osseiran, J. Monserrat, W. Mohr (Eds.), Mobile and Wireless Communications for IMT-Advanced and beyond, Wiley, 2011, pp. 121–155.
[41] W. Choi, J.G. Andrews, Downlink performance and capacity of distributed antenna systems in a multicell environment, IEEE Trans. Wireless Commun. 6 (1) (2007) 69–73.
[42] P. Marsch, G. Fettweis, On multicell cooperative transmission in backhaul-constrained cellular systems, Ann. Telecommun. 63 (2008) 253–269.
[43] J. Zhang, R. Chen, J.G. Andrews, A. Ghosh, R.W. Heath, Networked MIMO with clustered linear precoding, IEEE Trans. Wireless Commun. 8 (4) (2009) 1910–1921.
[44] H. Huang, M. Trivellato, A. Hottinen, M. Shafi, P.J. Smith, R. Valenzuela, Increasing downlink cellular throughput with limited network MIMO coordination, IEEE Trans. Wireless Commun. 8 (6) (2009) 2983–2989.
[45] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, L. Hanzo, Cell-free massive MIMO: a new next-generation paradigm, IEEE Access 7 (2019) 99878–99888.
[46] G. Interdonato, E. Björnson, H.Q. Ngo, P. Frenger, E.G. Larsson, Ubiquitous cell-free massive MIMO communications, EURASIP J. Wirel. Commun. Netw. 1 (2019) 197, 2019.
[47] T.L. Marzetta, E.G. Larsson, H. Yang, H.Q. Ngo, Fundamentals of Massive MIMO, Cambridge University Press, 2016.
[48] L. Sanguinetti, E. Björnson, J. Hoydis, Toward massive MIMO 2.0: understanding spatial correlation, interference suppression, and pilot contamination, IEEE Trans. Commun. 68 (1) (2019) 232–257.
[49] E. Björnson, L. Sanguinetti, Making cell-free massive MIMO competitive with MMSE processing and centralized implementation, IEEE Trans. Wireless Commun. 19 (1) (2019) 77–90.
[50] E. Nayebi, A. Ashikhmin, T.L. Marzetta, B.D. Rao, Performance of cell-free massive MIMO systems with MMSE and LSFD receivers, in: Proc. IEEE ACSSC, 2016, pp. 203–207.
[51] H.Q. Ngo, A. Ashikhmin, H. Yang, E.G. Larsson, T.L. Marzetta, Cell-free massive MIMO versus small cells, IEEE Trans. Wireless Commun. 16 (3) (2017) 1834–1850.
[52] E. Nayebi, A. Ashikhmin, T.L. Marzetta, H. Yang, B.D. Rao, Precoding and power optimization in cell-free massive MIMO systems, IEEE Trans. Wireless Commun. 16 (7) (2017) 4445–4459.
[53] Ö. Özdogan, E. Björnson, J. Zhang, Performance of cell-free massive MIMO with Rician fading and phase shifts, IEEE Trans. Wireless Commun. 18 (11) (2019) 5299–5315.
[54] Z. Wang, J. Zhang, E. Björnson, B. Ai, Uplink performance of cell-free massive MIMO over spatially correlated Rician fading channels, IEEE Commun. Lett. 25 (4) (2020) 1348–1352.
[55] S.-N. Jin, D.-W. Yue, H.H. Nguyen, Spectral and energy efficiency in cell-free massive MIMO systems over correlated Rician fading, IEEE Syst. J. 15 (2) (2021) 2822–2833.
[56] J. Zhang, J. Fan, B. Ai, D.W.K. Ng, NOMA-based cell-free massive MIMO over spatially correlated Rician fading channels, in: Proc. IEEE ICC, 2020, pp. 1–6.
[57] J. Zheng, J. Zhang, E. Björnson, B. Ai, Impact of channel aging on cell-free massive MIMO over spatially correlated channels, IEEE Trans. Wireless Commun. 20 (10) (2021) 6451–6466.
[58] E. Björnson, L. Sanguinetti, A new look at cell-free massive MIMO: making it practical with dynamic cooperation, in: Proc. IEEE PIMRC, 2019, pp. 1–6.
[59] A. Abdallah, M.M. Mansour, Efficient angle-domain processing for FDD-based cell-free massive MIMO systems, IEEE Trans. Commun. 68 (4) (2020) 2188–2203.
[60] S. Kim, J.W. Choi, B. Shim, Downlink pilot precoding and compressed channel feedback for FDD-based cell-free systems, IEEE Trans. Wireless Commun. 19 (6) (2020) 3658–3672.
[61] E. Björnson, N. Jalden, M. Bengtsson, B. Ottersten, Optimality properties, distributed strategies, and measurement-based evaluation of coordinated multicell OFDMA transmission, IEEE Trans. Signal Process. 59 (12) (2011) 6086–6101.
[62] E. Björnson, E. Jorswieck, Optimal resource allocation in coordinated multi-cell systems, Found. Trends® Commun. Inf. Theory 9 (2–3) (2013) 113–381.
[63] R. Nikbakht, A. Lozano, Uplink fractional power control for cell-free wireless networks, in: Proc. IEEE ICC, 2019, pp. 1–5.
[64] S. Chen, J. Zhang, E. Björnson, J. Zhang, B. Ai, Structured massive access for scalable cell-free massive MIMO systems, IEEE J. Sel. Area. Commun. 39 (4) (2020) 1086–1100.
[65] N. Shariati, E. Björnson, M. Bengtsson, M. Debbah, Low-complexity polynomial channel estimation in large-scale MIMO with arbitrary statistics, IEEE J. Sel. Top. Signal Process. 8 (5) (2014) 815–830.
[66] E. Bjrnson, P. Giselsson, Two applications of deep learning in the physical layer of communication systems [lecture notes], IEEE Signal Process. Mag. 37 (5) (2020) 134–140.
[67] Y. Jin, J. Zhang, S. Jin, B. Ai, Channel estimation for cell-free mmwave massive MIMO through deep learning, IEEE Trans. Veh. Technol. 68 (10) (2019) 10325–10329.
[68] Y. Jin, J. Zhang, B. Ai, X. Zhang, Channel estimation for mmwave massive MIMO with convolutional blind denoising network, IEEE Commun. Lett. 24 (1) (2019) 95–98.
[69] A. Almamori, S. Mohan, Estimation of channel state information (CSI) in cell-free massive MIMO based on time of arrival (ToA), Wireless Pers. Commun. 114 (2) (2020) 1825–1831.
[70] X. Huang, X. Zhu, Y. Jiang, Y. Liu, Efficient enhanced K-means clustering for semi-blind channel estimation of cell-free massive MIMO, in: Proc. IEEE ICC, 2020, pp. 1–6.
[71] A. Abdallah, M.M. Mansour, Angle-based multipath estimation and beamforming for FDD cell-free massive MIMO, in: Proc. IEEE SPAWC, 2019, pp. 1–5.
[72] A. Ashikhmin, H. Q. Ngo, T. L. Marzetta, H. Yang, Pilot assignment in cell free massive MIMO wireless systems, uS Patent 9,615,384 (Apr. 2017).
[73] T.C. Mai, H.Q. Ngo, T.Q. Duong, Uplink spectral efficiency of cell-free massive MIMO with multi-antenna users, in: Proc. IEEE SigTelCom, 2019, pp. 126–129.
[74] D. Maryopi, M. Bashar, A. Burr, On the uplink throughput of zero forcing in cell-free massive MIMO with coarse quantization, IEEE Trans. Veh. Technol. 68 (7) (2019) 7220–7224.

[75] M. Bashar, K. Cumanan, A.G. Burr, M. Debbah, H.Q. Ngo, On the uplink max–min SINR of cell-free massive MIMO systems, IEEE Trans. Wireless Commun. 18 (4) (2019) 2021–2036.
[76] G. Interdonato, M. Karlsson, E. Björnson, E.G. Larsson, Local partial zero-forcing precoding for cell-free massive MIMO, IEEE Trans. Wireless Commun. 19 (7) (2020) 4758–4774.
[77] F. Riera-Palou, G. Femenias, Trade-offs in cell-free massive MIMO networks: precoding, power allocation and scheduling, in: Proc. IEEE TELSIKS, 2019, pp. 158–165.
[78] M. Attarifar, A. Abbasfar, A. Lozano, Modified conjugate beamforming for cell-free massive MIMO, IEEE Wireless Commun. Lett. 8 (2) (2019) 616–619.
[79] G. Interdonato, M. Karlsson, E. Bjornson, E.G. Larsson, Downlink spectral efficiency of cell-free massive MIMO with full-pilot zero-forcing, in: Proc. IEEE GlobalSIP, 2018, pp. 1003–1007.
[80] X. Wang, M. Zhou, X. Qiao, Y. Zhang, X. Ruan, L. Yang, Energy efficiency of downlink cell-free massive MIMO system with full-pilot zero-forcing, in: Proc. IEEE ICCC, 2019, pp. 2065–2071.
[81] J. Zhang, J. Zhang, E. Björnson, B. Ai, Local partial zero-forcing combining for cell-free massive MIMO systems, IEEE Trans. Commun. Early Access 69 (12) (2021) 8459–8473.
[82] Ö. Özdogan, E. Björnson, J. Zhang, Downlink performance of cell-free massive MIMO with Rician fading and phase shifts, in: Proc. IEEE SPAWC, 2019, pp. 1–5.
[83] T.C. Mai, H.Q. Ngo, T.Q. Duong, Downlink spectral efficiency of cell-free massive MIMO systems with multi-antenna users, IEEE Trans. Commun. 68 (8) (2020) 4803–4815.
[84] G. Interdonato, H.Q. Ngo, P. Frenger, E.G. Larsson, Downlink training in cell-free massive MIMO: a blessing in disguise, IEEE Trans. Wireless Commun. 18 (11) (2019) 5153–5169.
[85] S. Buzzi, C. D'Andrea, C. D'Elia, User-centric cell-free massive MIMO with interference cancellation and local ZF downlink precoding, in: Proc. IEEE ISWCS, 2018, pp. 1–5.
[86] A. Zhou, J. Wu, E.G. Larsson, P. Fan, Max-min optimal beamforming for cell-free massive MIMO, IEEE Commun. Lett. 24 (10) (2020) 2344–2348.
[87] I. Atzeni, B. Gouda, A. Tölli, Distributed precoding design via over-the-air signaling for cell-free massive MIMO, IEEE Trans. Wireless Commun. 20 (2) (2020) 1201–1216.
[88] B. Gouda, I. Atzeni, A. Tölli, Distributed precoding design for cell-free massive MIMO systems, in: Proc. IEEE SPAWC, 2020, pp. 1–5.
[89] H.Q. Ngo, H. Tataria, M. Matthaiou, S. Jin, E.G. Larsson, On the performance of cell-free massive MIMO in ricean fading, in: Proc. IEEE ACSSC, 2018, pp. 980–984.
[90] T.K. Nguyen, T.H. Nguyen, Performance of assigning pilot sequences in cell free massive MIMO under SINR constraints, in: Proc. IEEE ICCE, 2018, pp. 121–126.
[91] Y. Zhang, H. Cao, P. Zhong, C. Qi, L. Yang, Location-based greedy pilot assignment for cell-free massive MIMO systems, in: Proc. IEEE ICCC, 2018, pp. 392–396.
[92] R. Sabbagh, C. Pan, J. Wang, Pilot allocation and sum-rate analysis in cell-free massive MIMO systems, in: Proc. IEEE ICC, 2018, pp. 1–6.
[93] H. Liu, J. Zhang, S. Jin, B. Ai, Graph coloring based pilot assignment for cell-free massive MIMO systems, IEEE Trans. Veh. Technol. 69 (8) (2020) 9180–9184.
[94] W. Hmida, V. Meghdadi, A. Bouallegue, J.P. Cances, Graph coloring based pilot reuse among interfering users in cell-free massive MIMO, in: Proc. IEEE ICC, 2020, pp. 1–6.
[95] H. Liu, J. Zhang, X. Zhang, A. Kurniawan, T. Juhana, B. Ai, Tabu-search-based pilot assignment for cell-free massive MIMO systems, IEEE Trans. Veh. Technol. 69 (2) (2019) 2286–2290.
[96] S. Buzzi, C. DAndrea, M. Fresia, Y.-P. Zhang, S. Feng, Pilot assignment in cell-free massive MIMO based on the Hungarian algorithm, IEEE Wireless Commun. Lett. 10 (1) (2020) 34–37.
[97] M. Attarifar, A. Abbasfar, A. Lozano, Random vs structured pilot assignment in cell-free massive MIMO wireless networks, in: Proc. IEEE ICC Workshops, 2018, pp. 1–6.
[98] U.K. Ganesan, E. Björnson, E.G. Larsson, An algorithm for grant-free random access in cell-free massive MIMO, in: Proc. IEEE SPAWC, 2020, pp. 1–5.
[99] M. Ke, Z. Gao, Y. Wu, X. Gao, K.-K. Wong, Massive access in cell-free massive MIMO-Based internet of things: cloud computing and edge computing paradigms, IEEE J. Sel. Area. Commun. 39 (3) (2020) 756–772.
[100] T. Van Chien, E. Björnson, E.G. Larsson, Joint power allocation and load balancing optimization for energy-efficient cell-free massive MIMO networks, IEEE Trans. Wireless Commun. 19 (10) (2020) 6798–6812.
[101] G. Femenias, N. Lassoued, F. Riera-Palou, Access point switch on ON/OFF strategies for green cell-free massive MIMO networking, IEEE Access 8 (2020) 21788–21803.
[102] J. García-Morales, G. Femenias, 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 8 (2020) 137587–137605.
[103] Y. Zhang, H. Cao, L. Yang, Max-min power optimization in multigroup multicast cell-free massive MIMO, in: Proc. IEEE WCNC, 2019, pp. 1–6.
[104] M. Bengtsson, B. Ottersten, Optimal and suboptimal transmit beamforming, in: L.C. Godara (Ed.), Handbook of Antennas in Wireless Communications, CRC Press, 2001.
[105] S. Buzzi, A. Zappone, Downlink power control in user-centric and cell-free massive MIMO wireless networks, in: Proc. IEEE PIMRC, 2017, pp. 1–6.
[106] Ö.T. Demir, E. Björnson, Joint power control and LSFD for wireless-powered cell-free massive MIMO, IEEE Trans. Wireless Commun. 20 (3) (2020) 1756–1769.
[107] M. Bashar, K. Cumanan, A.G. Burr, H.Q. Ngo, H.V. Poor, Mixed quality of service in cell-free massive MIMO, IEEE Commun. Lett. 22 (7) (2018) 1494–1497.
[108] M. Bashar, K. Cumanan, A.G. Burr, H.Q. Ngo, M. Debbah, P. Xiao, Max–min rate of cell-free massive MIMO uplink with optimal uniform quantization, IEEE Trans. Commun. 67 (10) (2019) 6796–6815.
[109] M. Bashar, K. Cumanan, A.G. Burr, M. Debbah, H.Q. Ngo, Enhanced max-min SINR for uplink cell-free massive MIMO systems, in: Proc. IEEE ICC, 2018, pp. 1–6.
[110] J. Zhang, J. Fan, J. Zhang, D.W.K. Ng, Q. Sun, B. Ai, Performance analysis and optimization of NOMA-based cell-free massive MIMO for IoT, IEEE Internet Things J. (2021) to appear.
[111] R. Nikbakht, R. Mosayebi, A. Lozano, Uplink fractional power control and downlink power allocation for cell-free networks, IEEE Wireless Commun. Lett. 9 (6) (2020) 774–777.
[112] Y. Zhao, I.G. Niemegeers, S.H. De Groot, Power allocation in cell-free massive MIMO: a deep learning method, IEEE Access 8 (2020) 87185–87200.
[113] S. Chakraborty, E. Björnson, L. Sanguinetti, Centralized and distributed power allocation for max-min fairness in cell-free massive MIMO, in: Proc. IEEE ACSSC, 2019, pp. 576–580.
[114] R. Nikbakht, A. Jonsson, A. Lozano, Unsupervised-learning power control for cell-free wireless systems, in: Proc. IEEE PIMRC, 2019, pp. 1–5.
[115] S.C. Tripathi, A. Trivedi, S. Rajoria, Power optimization of cell free massive MIMO with zero-forcing beamforming technique, in: Proc. IEEE CICT, 2018, pp. 1–4.
[116] T.H. Nguyen, T.K. Nguyen, H.D. Han, et al., Optimal power control and load balancing for uplink cell-free multi-user massive MIMO, IEEE Access 6 (2018) 14462–14473.
[117] J. Francis, P. Baracca, S. Wesemann, G. Fettweis, Downlink power control in cell-free massive MIMO with partially distributed access points, in: Proc. IEEE VTC-Fall, 2019, pp. 1–7.
[118] Y. Zhang, H. Cao, M. Zhou, L. Li, L. Yang, Power optimization in cell-free massive MIMO with non-ideal hardware transceiver, Chin. J. Electron. 29 (1) (2020) 190–198.
[119] Y. Zhang, H. Cao, Y. Guo, L. Yang, Sca power optimization in cell-free massive MIMO with short-term power constraints, in: Proc. IEEE WCSP, 2018, pp. 1–6.
[120] Y. Zhang, H. Cao, M. Zhou, S. Wu, L. Yang, Rate maximization for cell-free massive MIMO with low-resolution ADCs, in: Proc. IEEE DASC/PiCom/CBDCom/CyberSciTech, 2019, pp. 897–900.
[121] C. D'Andrea, A. Zappone, S. Buzzi, M. Debbah, Uplink power control in cell-free massive MIMO via deep learning, in: Proc. IEEE CAMSAP, 2019, pp. 554–558.
[122] M. Bashar, A. Akbari, K. Cumanan, H.-Q. Ngo, A.G. Burr, P. Xiao, M. Debbah, J. Kittler, Exploiting deep learning in limited-fronthaul cell-free massive MIMO uplink, IEEE J. Sel. Area. Commun. 38 (8) (2020) 1678–1697.
[123] F. Liang, C. Shen, W. Yu, F. Wu, Towards optimal power control via ensembling deep neural networks, IEEE Trans. Commun. 68 (3) (2019) 1760–1776.
[124] M. Alonzo, S. Buzzi, A. Zappone, C. D'Elia, Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave, IEEE Trans. Green Commun. Netw. 3 (3) (2019) 651–663.
[125] L.D. Nguyen, T.Q. Duong, H.Q. Ngo, K. Tourki, Energy efficiency in cell-free massive MIMO with zero-forcing precoding design, IEEE Commun. Lett. 21 (8) (2017) 1871–1874.
[126] M. Bashar, K. Cumanan, A.G. Burr, H.Q. Ngo, E.G. Larsson, P. Xiao, Energy efficiency of the cell-free massive MIMO uplink with optimal uniform quantization, IEEE Trans. Green Commun. Netw. 3 (4) (2019) 971–987.
[127] L.-N. Tran, H.Q. Ngo, First-order methods for energy-efficient power control in cell-free massive MIMO, in: Proc. IEEE ACSSC, 2019, pp. 848–852.
[128] H.Q. Ngo, L.-N. Tran, T.Q. Duong, M. Matthaiou, E.G. Larsson, On the total energy efficiency of cell-free massive MIMO, IEEE Trans. Green Commun. Netw. 2 (1) (2017) 25–39.
[129] J. Zhang, Y. Wei, E. Björnson, Y. Han, S. Jin, Performance analysis and power control of cell-free massive MIMO systems with hardware impairments, IEEE Access 6 (2018) 55302–55314.
[130] G. Interdonato, E. Björnson, H.Q. Ngo, P. Frenger, E.G. Larsson, Ubiquitous cell-free massive MIMO communications, EURASIP J. Wirel. Commun. Netw. (1) (2019) 197–209, 2019.
[131] E. Björnson, L. Sanguinetti, Cell-free versus cellular massive MIMO: what processing is needed for cell-free to win?, in: Proc. IEEE SPAWC, 2019, pp. 1–5.
[132] S. Buzzi, C. D'Andrea, A. Zappone, C. D'Elia, User-centric 5G cellular networks: resource allocation and comparison with the cell-free massive MIMO approach, IEEE Trans. Wireless Commun. 19 (2) (2019) 1250–1264.
[133] M. Bashar, A. Akbari, K. Cumanan, H.-Q. Ngo, A.G. Burr, P. Xiao, M. Debbah, J. Kittler, Exploiting deep learning in limited-fronthaul cell-free massive MIMO uplink, IEEE J. Sel. Area. Commun. 38 (8) (2020) 1678–1697.
[134] Ericsson, et al., Common Public Radio Interface (CPRI); Interface Specification V7. 0, Huawei Technologies Co. Ltd, NEC Corporation, Alcatel Lucent, and Nokia Networks.
[135] M. Bashar, K. Cumanan, A.G. Burr, M. Debbah, H.Q. Ngo, On the uplink max–min SINR of cell-free massive MIMO systems, IEEE Trans. Wireless Commun. 18 (4) (2019) 2021–2036.
[136] A. Burr, M. Bashar, D. Maryopi, Cooperative access networks: optimum fronthaul quantization in distributed massive mimo and cloud ran-invited paper, in: Proc. IEEE VTC Spring, 2018, pp. 1–5.
[137] M. Bashar, H.Q. Ngo, A.G. Burr, D. Maryopi, K. Cumanan, E.G. Larsson, On the performance of backhaul constrained cell-free massive MIMO with linear receivers, in: Proc. IEEE ACSSC, 2018, pp. 624–628.
[138] H. Masoumi, M.J. Emadi, Performance analysis of cell-free massive MIMO system with limited fronthaul capacity and hardware impairments, IEEE Trans. Wireless Commun. 19 (2) (2019) 1038–1053.
[139] H. Masoumi, M. J. Emadi, Transmission Strategies for Cell-free Massive MIMO with Limited Capacity Fronthaul Links, arXiv:1902.00700.

[140] M. Bashar, A. Akbari, K. Cumanan, H. Quoc Ngo, A.G. Burr, P. Xiao, M. Debbah, Deep Learning-Aided Finite-Capacity Fronthaul Cell-free Massive MIMO with Zero Forcing, 2020, pp. 1–6.
[141] G. Femenias, F. Riera-Palou, Fronthaul-constrained cell-free massive MIMO with low resolution ADCs, IEEE Access 8 (2020) 116195–116215.
[142] M. Bashar, K. Cumanan, A.G. Burr, H.Q. Ngo, M. Debbah, Cell-free massive mimo with limited backhaul, in: Proc. IEEE ICC, 2018, pp. 1–7.
[143] M. Bashar, K. Cumanan, A.G. Burr, H.Q. Ngo, E.G. Larsson, P. Xiao, On the energy efficiency of limited-backhaul cell-free massive MIMO, in: Proc. IEEE ICC, 2019, pp. 1–7.
[144] G. Femenias, F. Riera-Palou, Reduced-complexity downlink cell-free mmWave massive MIMO systems with fronthaul constraints, in: Proc. IEEE EUSIPCO, 2019, pp. 1–5.
[145] G. Femenias, F. Riera-Palou, Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity, IEEE Access 7 (2019) 44596–44612.
[146] P. Parida, H.S. Dhillon, A.F. Molisch, Downlink performance analysis of cell-free massive MIMO with finite fronthaul capacity, in: Proc. IEEE VTC-Fall, 2018, pp. 1–6.
[147] M.N. Boroujerdi, A. Abbasfar, M. Ghanbari, Cell free massive MIMO with limited capacity fronthaul, Wireless Pers. Commun. 104 (2) (2019) 633–648.
[148] Q. Huang, A. Burr, Compute-and-forward in cell-free massive mimo: great performance with low backhaul load, in: Proc. IEEE ICC Workshops, 2017, pp. 601–606.
[149] J. Zhang, J. Zhang, D.W.K. Ng, S. Jin, B. Ai, Improving sum-rate of cell-free massive MIMO with expanded compute-and-forward, IEEE Trans. Signal Process. (2022) to appear In press.
[150] B. Nazer, V.R. Cadambe, V. Ntranos, G. Caire, Expanding the compute-and-forward framework: unequal powers, signal levels, and multiple linear combinations, IEEE Trans. Inf. Theor. 62 (9) (2016) 4879–4909.
[151] J. Zheng, J. Zhang, L. Zhang, X. Zhang, B. Ai, Efficient receiver design for uplink cell-free massive MIMO with hardware impairments, IEEE Trans. Veh. Technol. 69 (4) (2020) 4537–4541.
[152] J. Zheng, J. Zhang, B. Ai, Uav communications with WPT-aided cell-free massive MIMO systems, IEEE J. Sel. Area. Commun. 39 (10) (2021) 3114–3128.
[153] X. Hu, C. Zhong, X. Chen, W. Xu, H. Lin, Z. Zhang, Cell-free massive MIMO systems with low resolution ADCs, IEEE Trans. Commun. 67 (10) (2019) 6844–6857.
[154] H. Masoumi, M.J. Emadi, Performance analysis of cell-free massive MIMO system with limited fronthaul capacity and hardware impairments, IEEE Trans. Wireless Commun. 19 (2) (2019) 1038–1053.
[155] X. Zhang, D. Guo, K. An, B. Zhang, Secure communications over cell-free massive MIMO networks with hardware impairments, IEEE Syst. J. 14 (2) (2019) 1909–1920.
[156] E. Björnson, M. Matthaiou, M. Debbah, Massive MIMO with non-ideal arbitrary arrays: hardware scaling laws and circuit-aware design, IEEE Trans. Wireless Commun. 14 (8) (2015) 4353–4368.
[157] G. Interdonato, P. Frenger, E.G. Larsson, Scalability aspects of cell-free massive MIMO, in: Proc. IEEE ICC, 2019, pp. 1–6.
[158] Z.H. Shaik, E. Björnson, E.G. Larsson, MMSE-optimal sequential processing for cell-free massive MIMO with radio stripes, IEEE Trans. Commun. 69 (11) (2021) 7775–7789.
[159] H.V. Balan, R. Rogalin, A. Michaloliakos, K. Psounis, G. Caire, Airsync: enabling distributed multiuser MIMO with full spatial multiplexing, IEEE/ACM Trans. Netw. 21 (6) (2013) 1681–1695.
[160] D. Wang, C. Zhang, Y. Du, J. Zhao, M. Jiang, X. You, Implementation of a cloud-based cell-free distributed massive MIMO system, IEEE Commun. Mag. 58 (8) (2020) 61–67.
[161] X. Xu, M. Tao, Modeling, analysis, and optimization of caching in multi-antenna small-cell networks, IEEE Trans. Wireless Commun. 18 (11) (2019) 5454–5469.
[162] S. Chen, J. Zhang, E. Björnson, S. Wang, C. Xing, B. Ai, Wireless caching: cell-free versus small cells, Proc. IEEE ICC (2021) 1–6.
[163] S. Mukherjee, J. Lee, Edge computing-enabled cell-free massive MIMO systems, IEEE Trans. Wireless Commun. 19 (4) (2020) 2884–2899.
[164] T.T. Vu, D.T. Ngo, N.H. Tran, H.Q. Ngo, M.N. Dao, R.H. Middleton, Cell-free massive MIMO for wireless federated learning, IEEE Trans. Wireless Commun. 19 (10) (2020) 6377–6392.
[165] T. T. Vu, D. T. Ngo, H. Q. Ngo, M. N. Dao, N. H. Tran, R. H. Middleton, User Selection Approaches to Mitigate the Straggler Effect for Federated Learning on Cell-free Massive MIMO Networks, arXiv:2009.02031.
[166] X. Li, E. Björnson, S. Zhou, J. Wang, in: Massive MIMO with Multi-Antenna Users: when Are Additional User Antennas Beneficial?, IEEE ICT, 2016, pp. 1–6.
[167] E. Björnson, M. Kountouris, M. Bengtsson, B. Ottersten, Receive combining vs. multi-stream multiplexing in downlink systems with multi-antenna users, IEEE Trans. Signal Process. 61 (13) (2013) 3431–3446.
[168] W. Jiang, H.D. Schotten, Deep learning for fading channel prediction, IEEE Open J. Commun. Soc. 1 (2020) 320–332.
[169] H. Kim, S. Kim, H. Lee, C. Jang, Y. Choi, J. Choi, Massive MIMO channel prediction: kalman filtering vs. machine learning, IEEE Trans. Commun. 69 (1) (2021) 518–528.
[170] F. Liu, C. Masouros, A. Li, H. Sun, L. Hanzo, MU-MIMO communications with MIMO radar: from co-existence to joint transmission, IEEE Trans. Wireless Commun. 17 (4) (2018) 2755–2770.
[171] F. Liu, C. Masouros, A.P. Petropulu, H. Griffiths, L. Hanzo, Joint radar and communication design: applications, state-of-the-art, and the road ahead, IEEE Trans. Commun. 68 (6) (2020) 3834–3862.