Making Cell-Free Massive MIMO Competitive With MMSE Processing and Centralized Implementation

Emil Björnson, *Senior Member, IEEE*, and Luca Sanguinetti, *Senior Member, IEEE*

***Abstract*— Cell-free Massive MIMO is considered as a promising technology for satisfying the increasing number of users and high rate expectations in beyond-5G networks. The key idea is to let many distributed access points (APs) communicate with all users in the network, possibly by using joint coherent signal processing. The aim of this paper is to provide the first comprehensive analysis of this technology under different degrees of cooperation among the APs. Particularly, the uplink spectral efficiencies of four different cell-free implementations are analyzed, with spatially correlated fading and arbitrary linear processing. It turns out that it is possible to outperform conventional Cellular Massive MIMO and small cell networks by a wide margin, but only using global or local minimum mean-square error (MMSE) combining. This is in sharp contrast to the existing literature, which advocates for maximum-ratio combining. Also, we show that a centralized implementation with optimal MMSE processing not only maximizes the SE but largely reduces the fronthaul signaling compared to the standard distributed approach. This makes it the preferred way to operate Cell-free Massive MIMO networks. Non-linear decoding is also investigated and shown to bring negligible improvements.**

***Index Terms*— Beyond 5G MIMO, cell-free massive MIMO, cellular massive MIMO, uplink, AP cooperation, MMSE processing, fronthaul signaling, non-linear decoding, small-cell networks.**

I. Introduction

THE traditional way to cover a large geographical area with wireless communication services uses the cellular network topology in Fig. 1(a), where each base station (BS) serves an exclusive set of user equipments (UEs) [2]. This network topology has been utilized for many decades and the spectral efficiency (SE) has been gradually improved by reducing the cell sizes and applying more advanced signal processing schemes for interference mitigation [3].

Recently, massive multiple-input multiple-output (mMIMO) has become the key 5G physical-layer technology [4]–[7]. It can improve the SE by at least $10\times$ over legacy cellular networks [3], by upgrading the BS hardware instead of deploying new BS sites. The SE gain comes from that each BS has a compact array with a hundred or more antennas, which are used for digital beamforming and, particularly, to spatially multiplex many user equipments (UEs) on the same time-frequency resource [8]. The characteristic feature of mMIMO, compared to traditional multi-user MIMO, is that each BS has many more antennas than UEs in the cell. Signal processing methods, such as minimum mean-squared error (MMSE) combining in the uplink, can be used individually at each BS to suppress interference from both the same and other cells [3], [9], [10], without the need for any BS cooperation. The mMIMO theory also supports deployments with spatially distributed arrays in each cell [11], [12], as also illustrated in Fig. 1(a). This setup is essentially the same as the Distributed Antenna System (DAS) setup in [13] and Coordinated Multi-Point (CoMP) with static, disjoint cooperation clusters [14], [15]. These are all different embodiments of cellular networks.

An alternative network infrastructure was considered in [16], [17] under the name of *Cell-free mMIMO*. The idea is to deploy a large number of distributed single-antenna access points (APs), which are connected to a central processing unit (CPU), also known as an edge-cloud processor [18] or cloud radio access network (C-RAN) data center [19]. The CPU operates the system in a Network MIMO fashion, with no cell boundaries, to jointly serve the UEs by coherent joint transmission and reception [15], [20]–[23]. Compared to traditional Network MIMO, the outstanding aspect of Cell-free mMIMO is the operating regime with many more APs than UEs [16]. From an analytical perspective, an important novelty was that imperfect channel state information (CSI) was considered in the performance analysis, while perfect CSI was often assumed in the past [15]. The paper [16] advocated the use of maximum ratio (MR) processing (a.k.a. matched filtering or conjugate beamforming) locally at each AP, while [17], [24] showed that partially or fully centralized processing at the CPU can achieve higher SE.

A. Motivation

The focus in the early papers [16], [17] was on comparing Cell-free mMIMO with a small-cell network; that is, the APs are deployed at the same places, but each AP serves its own exclusive set of UEs. Since small cells are a special case of Cell-free mMIMO, they obviously provide lower performance. Particularly, [16], [17] demonstrated large improvements in

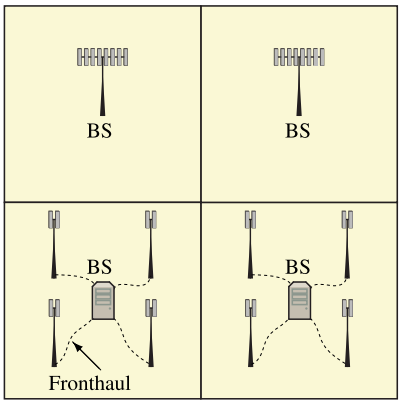
Manuscript received March 20, 2019; revised July 5, 2019 and September 3, 2019; accepted September 9, 2019. Date of publication September 20, 2019; date of current version January 8, 2020. The work of E. Björnson was supported by the Excellence Center at Linköping-Lund in Information Technology (ELLIIT) and the Wallenberg AI, Autonomous Systems and Software Program (WASP). The work of L. Sanguinetti was supported by the University of Pisa through the Research Project CONCEPT under Grant PRA 2018-2019. This article was presented in part at the IEEE SPAWC 2019 [1]. The associate editor coordinating the review of this article and approving it for publication was A. Zaidi. *(Corresponding author: Emil Björnson.)*

E. Björnson is with the Department of Electrical Engineering (ISY), Linköping University, 58183 Linköping, Sweden (e-mail: emil.bjornson@liu.se).

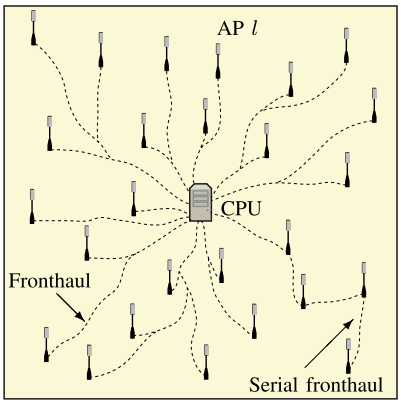
L. Sanguinetti is with the Dipartimento di Ingegneria dell'Informazione, University of Pisa, 56122 Pisa, Italy (e-mail: luca.sanguinetti@unipi.it).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2941478



(a) Cellular network with mMIMO BSs having either co-located arrays (top) or distributed arrays (bottom).



(b) Cell-free mMIMO network.

Fig. 1. Comparison of different cellular and cell-free network topologies.

median and 95%-likely SE. In Section IV, we will show that this is partially due to the fact that a poor implementation of the small-cell network was considered in [16], [17]. In fact, we will show that more sophisticated processing than MR is needed in Cell-free mMIMO to always outperform small cells.

Unlike [16], [17], this paper aims at comparing Cell-free mMIMO with conventional Cellular mMIMO and its primary goal is to find the most competitive cell-free implementation.[1] Both network topologies are illustrated in Fig. 1. The large differences make the comparison non-trivial and provide interesting inputs into the design of beyond-5G networks. Cellular mMIMO benefits from channel hardening and spatial interference suppression, but cell-edge UEs can have bad channel conditions. On the other hand, Cell-free mMIMO benefits from strong macro diversity but its interference suppression capability highly depends on how it is operated. The early papers [16], [17] conjectured that channel hardening also appears in Cell-free mMIMO, but it was later shown that capacity bounds that presumes hardening can greatly underestimate the practical SE [27]. To achieve a reasonably fair comparison, we focus on the uplink and assume that the data transmission is preceded by a pilot-based channel estimation phase. All UEs transmit with equal powers for any of the different levels and network topologies.

[1]Previous comparisons are found in [25], [26] but only for a single cell, so it is not cellular, and only MR is used, which is known to perform badly [3].

B. Contributions

The major contributions of this paper are two-fold. Firstly, we introduce a taxonomy with four different implementations of Cell-free mMIMO, which are characterized by different degrees of cooperation among the APs. Secondly, we provide new achievable SE expressions, which are valid for spatially correlated fading channels, imperfect CSI, APs with an arbitrary number N of antennas, and heuristic or optimized receive combining schemes. All this provides a common analytical framework to numerically evaluate the benefits and costs (in terms of fronthaul signaling) of the different implementations and to understand how Cell-free mMIMO should be operated and designed in order to get much higher performance than conventional Cellular mMIMO and small cells.

The four different levels of cooperation that we consider in this paper are as follows. The so-called *Level 4* is a form of Network MIMO and stands for a fully centralized network in which the pilot and data signals received at all APs are gathered (through the fronthaul links) at the CPU, which performs channel estimation and data detection. *Level 3* relies on the large-scale fading decoding (LSFD) strategy, which was originally proposed for Cellular mMIMO in [28], [29]. Particularly, it operates in two stages. In the first stage, each AP locally estimates the channels and applies an arbitrary receive combiner to obtain local estimates of the UE data. These are then gathered at the CPU where they are linearly processed to perform joint detection. Only channel statistics can be utilized in the second stage at the CPU since the pilot signals are not shared over the fronthaul links. *Level 2* is a direct simplification of Level 3 in the sense that the CPU performs detection in the second stage by simply taking the average of the local estimates. This dispenses the CPU from knowledge of the channel statistics and thus reduces the amount of information to be exchanged. Finally, *Level 1* stands for a fully distributed network in which the detection is performed locally at the APs by using only local channel estimates and one AP serves each UE. This is a small-cell network where nothing is exchanged with the CPU.

The above levels have been partially analyzed before in the literature, but not under the general and practical conditions considered in this paper, which allow us to draw conclusions that differ in several important ways—in particular, we show that MR combining performs terribly bad in Cell-free mMIMO. Level 4 was considered in [24], [30], [31] for $N = 1$ and in [27], [32] with $N \geq 1$ but with spatially uncorrelated channels. Level 3 was investigated in [24], [33] for $N = 1$ and MR combining. Level 2 was considered in [16], [31], [34]–[38] (among many others) but only with MR combining. A suboptimal implementation of Level 1 with $N = 1$ was considered in [16] (the suboptimality is explained in detail in Section III-D). There are also previous papers that consider various forms of Levels 1–4 under perfect CSI; see the reference list of [16] for a good selection of

such papers. In addition, there is previous research on BS cooperation in cellular networks, where the received signals and CSI are shared between BSs to cancel inter-cell interference; see [39]–[41] and reference therein. These papers also consider different levels of cooperation, but these are heavily influenced by the cellular topology (e.g., BSs send signals to each other, BSs are surrounded by UEs, and there exist cell edges) and can thus not be applied to Cell-free mMIMO.

C. Paper Outline

The rest of this paper is organized as follows. Section II defines the system model for uplink Cell-free mMIMO for both data transmission and channel estimation. Next, Section III presents the four levels of receiver cooperation, including achievable SE expressions for spatially correlated fading, multi-antenna APs, and optimized receive combining. The four levels are numerically compared with Cellular mMIMO in Section IV. This section also discusses the differences and similarities with the previous results in [16]. Section V evaluates the potential benefit of using non-linear decoding at the CPU, whereas the fronthaul signaling required with the different implementations is quantified in Section VI. Finally, the major conclusions and implications are drawn in Section VII.

Reproducible research: All the simulation results can be reproduced using the Matlab code and data files available at: https://github.com/emilbjornson/competitive-cell-free

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 $\mathbb{E}\{\mathbf{x}\}$.

II. Cell-Free Network Model

We consider a Cell-free mMIMO network consisting of L geographically distributed APs, each equipped with N antennas. The APs are connected via fronthaul connections to a CPU, as illustrated in Fig. 1(b). There are K single-antenna UEs in the network and the channel between AP l and UE k is denoted by $\mathbf{h}_{kl} \in \mathbb{C}^N$. We use the standard block fading model where $\mathbf{h}_{kl}$ is constant in time-frequency blocks of τ_c channel uses [3]. In each block, an independent realization from a correlated Rayleigh fading distribution is drawn:

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

where $\mathbf{R}_{kl} \in \mathbb{C}^{N \times N}$ is the spatial correlation matrix, which describes the spatial properties of the channel and $\beta_{kl} \triangleq \mathrm{tr}(\mathbf{R}_{kl})/N$ is the large-scale fading coefficient that describes geometric pathloss and shadowing.

This paper considers the uplink, which consists of τ_p channel uses dedicated for pilots and $\tau_c - \tau_p$ channel uses for payload data. The two phases are described below. Notice that the results of this paper apply to both systems operating in time-division duplex (TDD) and frequency-division duplex (FDD) mode, since the uplink works the same in both cases.

A. Pilot Transmission and Channel Estimation

We assume that τ_p mutually orthogonal τ_p-length pilot signals $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{\tau_p}$ with $\|\boldsymbol{\phi}_t\|^2 = \tau_p$ are used for channel estimation. These pilots are assigned to the UEs in a deterministic but arbitrary way. The case of practical interest is a large network with $K > \tau_p$ so that more than one UE is assigned to each pilot. We denote the index of the pilot assigned to UE k as $t_k \in \{1, \ldots, \tau_p\}$ and call $\mathcal{P}_k \subset \{1, \ldots, K\}$ the subset of UEs that use the same pilot as UE k, including itself.

When the UEs transmit their pilots, the received signal $\mathbf{Z}_l \in \mathbb{C}^{N \times \tau_p}$ at AP l is

$$\mathbf{Z}_l = \sum_{i=1}^{K} \sqrt{p_i} \mathbf{h}_{il} \boldsymbol{\phi}_{t_i}^{\mathrm{T}} + \mathbf{N}_l \tag{2}$$

where $p_i \geq 0$ is the transmit power of UE i, $\mathbf{N}_l \in \mathbb{C}^{N \times \tau_p}$ is the receiver noise with independent $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ entries, and σ^2 is the noise power. To estimate $\mathbf{h}_{kl}$, the AP first correlates the received signal with the associated normalized pilot signal $\boldsymbol{\phi}_{t_k}/\sqrt{\tau_p}$ to obtain $\mathbf{z}_{t_k l} \triangleq \frac{1}{\sqrt{\tau_p}} \mathbf{Z}_l \boldsymbol{\phi}_{t_k}^* \in \mathbb{C}^N$, which is given by

$$\begin{aligned} \mathbf{z}_{t_k l} &= \sum_{i=1}^{K} \frac{\sqrt{p_i}}{\sqrt{\tau_p}} \mathbf{h}_{il} \boldsymbol{\phi}_{t_i}^{\mathrm{T}} \boldsymbol{\phi}_{t_k}^* + \frac{1}{\sqrt{\tau_p}} \mathbf{N}_l \boldsymbol{\phi}_{t_k}^* \\ &= \sum_{i \in \mathcal{P}_k} \sqrt{p_i \tau_p} \mathbf{h}_{il} + \mathbf{n}_{t_k l} \end{aligned} \tag{3}$$

where $\mathbf{n}_{t_k l} \triangleq \mathbf{N}_l \boldsymbol{\phi}_{t_k}^* / \sqrt{\tau_p} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ is the resulting noise. Using standard results from estimation theory [3, Sec. 3], the MMSE estimate of $\mathbf{h}_{kl}$ is

$$\hat{\mathbf{h}}_{kl} = \sqrt{p_k \tau_p} \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{z}_{t_k l} \tag{4}$$

where

$$\boldsymbol{\Psi}_{t_k l} = \mathbb{E}\{\mathbf{z}_{t_k l} \mathbf{z}_{t_k l}^{\mathrm{H}}\} = \sum_{i \in \mathcal{P}_k} \tau_p p_i \mathbf{R}_{il} + \mathbf{I}_N \tag{5}$$

is the correlation matrix of the received signal in (3). The estimate $\hat{\mathbf{h}}_{kl}$ and estimation error $\tilde{\mathbf{h}}_{kl} = \mathbf{h}_{kl} - \hat{\mathbf{h}}_{kl}$ are independent vectors distributed as $\hat{\mathbf{h}}_{kl} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, p_k \tau_p \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{R}_{kl}\right)$ and $\tilde{\mathbf{h}}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{C}_{kl})$ with

$$\mathbf{C}_{kl} = \mathbb{E}\{\tilde{\mathbf{h}}_{kl} \tilde{\mathbf{h}}_{kl}^{\mathrm{H}}\} = \mathbf{R}_{kl} - p_k \tau_p \mathbf{R}_{kl} \boldsymbol{\Psi}_{t_k l}^{-1} \mathbf{R}_{kl}. \tag{6}$$

The mutual interference generated by the pilot-sharing UEs in (3) causes the so-called *pilot contamination* that degrades the system performance, similar to the case in Cellular mMIMO.

Remark 1*: The computation of* $\hat{\mathbf{h}}_{kl}$ *in* (4) *requires knowledge of the correlation matrices* $\{\mathbf{R}_{il} : i \in \mathcal{P}_k\}$*, which we assume to be locally available at AP* l*; see [3] for methods to estimate them. To dispense with their full knowledge, the AP can apply alternative channel estimation schemes as in Cellular mMIMO [3, Sec. 3.4]. One option is the so-called element-wise MMSE estimator that uses only the main diagonals of*

$\{\mathbf{R}_{il} : i \in \mathcal{P}_k\}$. *Alternatively, the least-square estimator can be used, which requires no prior statistical information and computes the estimate of* $\mathbf{h}_{kl}$ *as* $\hat{\mathbf{h}}_{kl} = \frac{1}{\sqrt{p_k \tau_p}} \mathbf{z}_{t_k l}$*; see [37].*

B. Uplink Data Transmission

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

$$\mathbf{y}_l = \sum_{i=1}^{K} \mathbf{h}_{il} s_i + \mathbf{n}_l \tag{7}$$

where $s_i \sim \mathcal{N}_{\mathbb{C}}(0, p_i)$ is the information-bearing signal transmitted by UE i with power p_i and $\mathbf{n}_l \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ is the independent receiver noise.

Remark 2*: The signal model in* (2) *and* (7) *implicitly assumes that the entire network is synchronized in time. There exist wired and over-the-air methods that can be used to synchronize the clocks at the APs [19], [42], [43]. However, the signal transmitted by a UE will never be synchronously received by all the APs due to the largely different distances between the UE and different APs. In orthogonal frequency-division multiplexing systems, a simple way to compensate for that is to select the length of the cyclic prefix so as to accommodate both the channel delay spread and timing misalignments. This results in a quasi-synchronous system [44]. For example, in the LTE standard, the cyclic prefix is long enough to assume that a UE is quasi-synchronized to all APs within a 1 km radius. If the extended cyclic prefix is used, the range increases up to 5 km. Since the APs that are further away will receive negligible signal power, the model in* (2) *and* (7) *is accurate enough for the performance analysis considered in this paper.*

III. Four Levels of Receiver Cooperation

All the APs are connected via fronthaul connections to a CPU that has high computational resources.[2] Hence, the APs can be viewed as remote-radio heads that cooperate to support coherent communication with the UEs. The fronthaul can consist of a mix of wired and wireless connections, organized in a star or mesh topology [19]; the methods developed in this paper can be applied with any fronthaul topology. AP l receives the signal $\mathbf{y}_l$ in (7) and can use the available channel estimates $\{\hat{\mathbf{h}}_{kl} : k = 1, \ldots, K\}$ to detect the data signals locally, or can fully or partially delegate this task to the CPU. The benefit of using the CPU is that it can combine the inputs from all APs, but this must be balanced against the required amount of fronthaul signaling. Four levels of receiver cooperation are described below and compared with Cellular mMIMO in Section IV by means of numerical results.

A. Level 4: Fully Centralized Processing

The most advanced level of Cell-free mMIMO operation is when the L APs send their received pilot signals $\{\mathbf{z}_{tl} : t = 1, \ldots, \tau_p, l = 1, \ldots, L\}$ and received data signals $\{\mathbf{y}_l : l = 1, \ldots, L\}$ to the CPU, which takes care of the channel estimation and data signal detection. In other words, the APs act as relays that forward all signals to the CPU [45]. In each coherence block, each AP needs to send $\tau_p N$ complex scalars for the pilot signals and $(\tau_c - \tau_p)N$ complex scalars for the received signals. This becomes $\tau_c N$ complex scalars in total, which is summarized in Table I. Moreover, the spatial correlation matrices $\{\mathbf{R}_{kl} : k = 1, \ldots, K, l = 1, \ldots, L\}$ are assumed available at the CPU at Level 4, which are described by KLN^2 real scalars or $KLN^2/2$ complex scalars.[3]

TABLE I
Number of Complex Scalars to Send From the APs to the CPU via the Fronthaul, Either in Each Coherence Block or for Each Realization of the User Locations/Statistics.

	Each coherence block	Statistical parameters
Level 4	$\tau_c NL$	$KLN^2/2$
Level 3	$(\tau_c - \tau_p)KL$	$KL + (L^2K^2 + KL)/2$
Level 2	$(\tau_c - \tau_p)KL$	–
Level 1	–	–

The received signal available at the CPU is expressed as

$$\underbrace{\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_L \end{bmatrix}}_{\triangleq \mathbf{y}} = \sum_{i=1}^{K} \underbrace{\begin{bmatrix} \mathbf{h}_{i1} \\ \vdots \\ \mathbf{h}_{iL} \end{bmatrix}}_{\triangleq \mathbf{h}_i} s_i + \underbrace{\begin{bmatrix} \mathbf{n}_1 \\ \vdots \\ \mathbf{n}_L \end{bmatrix}}_{\triangleq \mathbf{n}} \tag{8}$$

or, in a more compact form, as

$$\mathbf{y} = \sum_{i=1}^{K} \mathbf{h}_i s_i + \mathbf{n}. \tag{9}$$

The collective channel is distributed as $\mathbf{h}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{R}_k)$ where $\mathbf{R}_k = \mathrm{diag}(\mathbf{R}_{k1}, \ldots, \mathbf{R}_{kL}) \in \mathbb{C}^{LN \times LN}$ is the block-diagonal spatial correlation matrix. Notice that (9) is mathematically equivalent to the signal model of a single-cell mMIMO system with correlated fading [3, Sec. 2.3.1]. The only difference is how the correlation matrices are generated and how the pilots are allocated. In fact, in conventional single-cell mMIMO orthogonal pilots are assigned to UEs whereas the same pilot can be assigned to multiple UEs in the investigated cell-free network. This leads to pilot contamination between UEs served by the same AP antennas.

The CPU can compute all the MMSE channel estimates $\{\hat{\mathbf{h}}_{kl} : k = 1, \ldots, K, l = 1, \ldots, L\}$ using the received pilot signals and channel statistics obtained from the APs. The estimates can be computed separately without loss of optimality. For UE k, the CPU can then form the collective channel estimate

$$\hat{\mathbf{h}}_k \triangleq \begin{bmatrix} \hat{\mathbf{h}}_{k1} \\ \vdots \\ \hat{\mathbf{h}}_{kL} \end{bmatrix} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, p_k \tau_p \mathbf{R}_k \boldsymbol{\Psi}_{t_k}^{-1} \mathbf{R}_k \right) \tag{10}$$

[2] In practice, cell-free systems will have more than one CPU and only a subset of the APs will serve each UE [15], [19], [43]. The methods described in this paper applies also to that case. The only requirement is that each UE is assigned to one CPU that takes partial or full responsibility for the decoding of the UE's data and will then forward the decoded data to the core network.

[3] It is not strictly necessary for the CPU to know the spatial correlation matrices, but it can use estimators that do not require that; see Remark 1.

where $\boldsymbol{\Psi}_{t_k}^{-1} = \mathrm{diag}(\boldsymbol{\Psi}_{t_k1}^{-1}, \ldots, \boldsymbol{\Psi}_{t_kL}^{-1})$. The estimation error is $\tilde{\mathbf{h}}_k = \mathbf{h}_k - \hat{\mathbf{h}}_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{C}_k)$ with $\mathbf{C}_k = \mathrm{diag}(\mathbf{C}_{k1}, \ldots, \mathbf{C}_{kL})$. Next, the CPU selects an arbitrary receive combining vector $\mathbf{v}_k \in \mathbb{C}^{LN}$ for UE k based on all the collective channel estimates $\{\hat{\mathbf{h}}_k : k = 1, \ldots, K\}$.

While the capacity of Level 4 networks with perfect CSI is known in some cases [45], the ergodic capacity is generally unknown in the considered case with imperfect CSI. However, we can rigorously analyze the performance by using standard capacity lower bounds [3], [46], which we refer to as achievable SEs.

Proposition 1*: At Level 4, if the MMSE estimator is used to compute channel estimates for all UEs, an achievable SE of UE k is*

$$\mathsf{SE}_k^{(4)} = \left(1 - \frac{\tau_p}{\tau_c}\right) \mathbb{E}\left\{\log_2\left(1 + \mathsf{SINR}_k^{(4)}\right)\right\} \tag{11}$$

where the instantaneous effective signal-to-interference-and-noise ratio (SINR) is

$$\mathsf{SINR}_k^{(4)} = \frac{p_k |\mathbf{v}_k^{\mathrm{H}} \hat{\mathbf{h}}_k|^2}{\sum\limits_{i=1, i\neq k}^{K} p_i |\mathbf{v}_k^{\mathrm{H}} \hat{\mathbf{h}}_i|^2 + \mathbf{v}_k^{\mathrm{H}} \left(\sum\limits_{i=1}^{K} p_i \mathbf{C}_i + \sigma^2 \mathbf{I}_{LN}\right) \mathbf{v}_k} \tag{12}$$

and the expectation is with respect to the channel estimates.

Proof: The proof follows the same steps as the proof of [3, Th. 4.1] for Cellular mMIMO and is therefore omitted. ■

The pre-log factor $1 - \tau_p/\tau_c$ in (11) is the fraction of channel uses that are used for uplink data transmission. The term $\mathsf{SINR}_k^{(4)}$ takes the form of an "effective instantaneous SINR" [3], with the desired signal power received over the estimated channel in the numerator and the interference plus noise in the denominator.[4]

We notice that the SE expression in (11) holds for any receive combining vector $\mathbf{v}_k$ and is a multi-antenna generalization of [24, Eq. (1)] and an extension of [27], [32] to spatially correlated channels. The expression can be easily computed for any $\mathbf{v}_k$ by using Monte Carlo methods, as done in Section IV. A possible choice is to use the simple MR combining with $\mathbf{v}_k = \hat{\mathbf{h}}_k$, which has low computational complexity and maximizes the power of the desired signal, but neglects the existence of interference. Other heuristic combiners such as zero-forcing (ZF) or regularized zero-forcing (RZF) can be also applied. Instead of resorting to heuristics, we notice that $\mathsf{SINR}_k^{(4)}$ in (12) only depends on $\mathbf{v}_k$ and has the form of a generalized Rayleigh quotient. Hence, the combining vector that maximizes (12) can be obtained as follows.

Corollary 1*: The instantaneous SINR in* (12) *for UE k is maximized by the MMSE combining vector*

$$\mathbf{v}_k = p_k \left(\sum_{i=1}^{K} p_i \left(\hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^{\mathrm{H}} + \mathbf{C}_i\right) + \sigma^2 \mathbf{I}_{LN}\right)^{-1} \hat{\mathbf{h}}_k \tag{13}$$

which leads to the maximum value

$$\mathsf{SINR}_k^{(4)} = p_k \hat{\mathbf{h}}_k^{\mathrm{H}} \left(\sum_{i=1, i\neq k}^{K} p_i \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^{\mathrm{H}} + \sum_{i=1}^{K} p_i \mathbf{C}_i + \sigma^2 \mathbf{I}_{LN}\right)^{-1} \hat{\mathbf{h}}_k. \tag{14}$$

Proof: It follows from [3, Lemma B.10] since (11) is a generalized Rayleigh quotient with respect to $\mathbf{v}_k$. ■

It can be shown that the SINR-maximizing combiner in (13) minimizes the mean-squared error $\mathrm{MSE}_k = \mathbb{E}\{|s_k - \mathbf{v}_k^{\mathrm{H}} \mathbf{y}|^2 \big| \{\hat{\mathbf{h}}_i\}\}$, which represents the conditional MSE between the data signal s_k and the received signal $\mathbf{v}_k^{\mathrm{H}} \mathbf{y}$ after receive combining; see [3, Sec. 4.1] for details. This is why it is called *MMSE combining*. This type of receive combining normally maximizes the mutual information of channels with multiple receive antennas [47], but the particular expression in (13) is unique for Cell-free mMIMO.

Compared to many heuristic solutions, MMSE combining has higher computational complexity since it requires first the computation of the $LN \times LN$ matrix inverse in (13) and then a matrix-vector multiplication. However, this is not a major issue since it has to be implemented at the CPU, which is assumed to have high computational capability. If the complexity is a concern, then ZF and RZF can be used instead since only $K \times K$ matrices need to be inverted. The price to pay is that the UEs with low SNRs get an SE reduction, which may be very large.

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

Instead of sending the N-dimensional vectors $\{\mathbf{y}_l : l = 1, \ldots, L\}$ and the channel estimates to the CPU, each AP can preprocess its signal by computing local estimates of the data that are then passed to the CPU for final decoding. Let $\mathbf{v}_{kl} \in \mathbb{C}^N$ be the local combining vector that AP l selects for UE k. Then, its local estimate of s_k is

$$\check{s}_{kl} \triangleq \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{y}_l = \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{h}_{kl} s_k + \sum_{i=1, i\neq k}^{K} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{h}_{il} s_i + \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{n}_l. \tag{15}$$

Any combining vector can be adopted in the above expression. Unlike at Level 4, however, AP l can only use its own local channel estimates $\{\hat{\mathbf{h}}_{il} : i = 1, \ldots, K\}$ for the design of $\mathbf{v}_{kl}$. The simplest solution is MR combining with $\mathbf{v}_{kl} = \hat{\mathbf{h}}_{kl}$ as in [16], [24] but preferably the AP should use its local CSI to make $\check{s}_{kl}$ as close to s_k as possible. The combining vector that minimizes the MSE, $\mathrm{MSE}_{kl} = \mathbb{E}\{|s_k - \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{y}_l|^2 \big| \{\hat{\mathbf{h}}_{il}\}\}$, is

$$\mathbf{v}_{kl} = p_k \left(\sum_{i=1}^{K} p_i \left(\hat{\mathbf{h}}_{il} \hat{\mathbf{h}}_{il}^{\mathrm{H}} + \mathbf{C}_{il}\right) + \sigma^2 \mathbf{I}_N\right)^{-1} \hat{\mathbf{h}}_{kl} \tag{16}$$

which can be proved by computing the conditional expectation and equating the first derivative with respect to $\mathbf{v}_{kl}$ to zero. Notice that (16) is the combining vector that would maximize the SE if AP l decoded the data signal s_k locally. We call (16) *Local MMSE (L-MMSE) combining* to distinguish it from the MMSE combining in (13) at Level 4, which is

[4]The word "effective" refers to the fact that $\mathsf{SINR}_k^{(4)}$ cannot be measured in the system at any particular point in time, but the SE is the same as that of a fading single-antenna point-to-point channel where $\mathsf{SINR}_k^{(4)}$ is the instantaneously measurable SINR and the receiver has perfect CSI.

applied at the CPU. A main benefit over MMSE combining is that an $N \times N$ matrix is inverted in (16) instead of an $LN \times LN$ matrix. Importantly, even if $N = 1$, (16) is not equal to MR but differ by a non-deterministic scaling factor.

The local estimates $\{\check{s}_{kl} : l = 1, \ldots, L\}$ are then sent to the CPU where they are linearly combined using the weights $\{a_{kl} : l = 1, \ldots, L\}$ to obtain $\hat{s}_k = \sum_{l=1}^{L} a_{kl}^* \check{s}_{kl}$, which is eventually used to decode s_k. From (15), we have that

$$\hat{s}_k = \left(\sum_{l=1}^{L} a_{kl}^* \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{h}_{kl} \right) s_k + \sum_{l=1}^{L} a_{kl}^* \left(\sum_{i=1, i \neq k}^{K} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{h}_{il} s_i \right) + \mathbf{n}_k' \tag{17}$$

with $\mathbf{n}_k' = \sum_{l=1}^{L} a_{kl}^* \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{n}_l$. Let $\mathbf{g}_{ki} = [\mathbf{v}_{k1}^{\mathrm{H}} \mathbf{h}_{i1} \ldots \mathbf{v}_{kL}^{\mathrm{H}} \mathbf{h}_{iL}]^{\mathrm{T}}$ be the L-dimensional vector with the receive-combined channels between UE k and each of the APs. Then, (17) reduces to

$$\hat{s}_k = \mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk} s_k + \sum_{i=1, i \neq k}^{K} \mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki} s_i + \mathbf{n}_k' \tag{18}$$

where $\mathbf{a}_k = [a_{k1} \ \ldots \ a_{kL}]^{\mathrm{T}} \in \mathbb{C}^L$ is the weighting coefficient vector and $\{\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki} : i = 1, \ldots, K\}$ represent the effective channels. Notice that $\mathbf{a}_k$ can be optimized by the CPU to maximize the SE, but only channel statistics can be utilized since the CPU does not have knowledge of the channel estimates at Level 3. This approach is known as LSFD in Cellular mMIMO [28], [29], and can be applied at Level 3 as follows. Although the effective channel $\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk}$ is unknown at the CPU, we notice that its average $\mathbf{a}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\}$ is non-zero (e.g., if L-MMSE or MR is used) and deterministic. Therefore, it can be assumed known[5] and used to compute the following achievable SE.

Proposition 2*: At Level 3, an achievable SE of UE k is*

$$\mathsf{SE}_k^{(3)} = \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \mathsf{SINR}_k^{(3)}\right) \tag{19}$$

with the effective SINR given by

$$\mathsf{SINR}_k^{(3)} = \frac{p_k \left| \mathbf{a}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\} \right|^2}{\sum_{i=1}^{K} p_i \mathbb{E}\{|\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{ki}|^2\} - p_k \left| \mathbf{a}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\} \right|^2 + \sigma^2 \mathbf{a}_k^{\mathrm{H}} \mathbf{D}_k \mathbf{a}_k} \tag{20}$$

where $\mathbf{D}_k = \mathrm{diag}(\mathbb{E}\{\|\mathbf{v}_{k1}\|^2\}, \ldots, \mathbb{E}\{\|\mathbf{v}_{kL}\|^2\}) \in \mathbb{C}^{L \times L}$ *and the expectations are with respect to all sources of randomness.*

Proof: The proof is given in Appendix A. ∎

The achievable SE above holds for any combining scheme. Particularly, it is valid for both the L-MMSE combining in (16) and the MR combining $\mathbf{v}_{kl} = \hat{\mathbf{h}}_{kl}$ that was used in [24]. Unlike the achievable SE in Proposition 1, it holds for any channel estimator (not only for the MMSE estimator (11)) but requires channel hardening in order to approximate $\mathbf{a}_k^{\mathrm{H}} \mathbf{g}_{kk}$ with its mean value $\mathbf{a}_k^{\mathrm{H}} \mathbb{E}\{\mathbf{g}_{kk}\}$. However, this may not occur when the number N of antennas at the APs is relatively small [27]. In that case, the SE expression in (20) underestimates the achievable performance, but is anyway the best available capacity bound.

The structure of (20) allows computing the deterministic weighting vector $\mathbf{a}_k$ that maximizes $\mathsf{SINR}_k^{(3)}$. This is given as follows.

Corollary 2*: The effective SINR in* (20) *for UE k is maximized by*

$$\mathbf{a}_k = \left(\sum_{i=1}^{K} p_i \mathbb{E}\{\mathbf{g}_{ki} \mathbf{g}_{ki}^{\mathrm{H}}\} + \sigma^2 \mathbf{D}_k \right)^{-1} \mathbb{E}\{\mathbf{g}_{kk}\} \tag{21}$$

which leads to the maximum value

$$\mathsf{SINR}_k^{(3)} = p_k \mathbb{E}\{\mathbf{g}_{kk}^{\mathrm{H}}\} \left(\sum_{i=1}^{K} p_i \mathbb{E}\{\mathbf{g}_{ki} \mathbf{g}_{ki}^{\mathrm{H}}\} + \sigma^2 \mathbf{D}_k - p_k \mathbb{E}\{\mathbf{g}_{kk}\} \mathbb{E}\{\mathbf{g}_{kk}^{\mathrm{H}}\} \right)^{-1} \mathbb{E}\{\mathbf{g}_{kk}\}. \tag{22}$$

Proof: It follows from [3, Lemma B.10] by noting that (20) is a generalized Rayleigh quotient with respect to $\mathbf{a}_k$. ∎

Notice that Level 3 is an extension of the LSFD framework in [24], [29], [33], [48], which has previously been only used in Cell-free mMIMO along with MR combining. In fact, the SE expressions provided in these papers only apply for particular choices of receive combining and not for arbitrary combining as (19). This makes Proposition 2 a novel contribution of this paper.

The signaling required at Level 3 can be quantified as follows. Each AP needs to send $(\tau_c - \tau_p)K$ complex scalars (i.e., $\check{s}_{kl}$ for all k) to the CPU per coherence block. In addition, the computation of (21) requires knowledge of the L-dimensional complex vector $\mathbb{E}\{\mathbf{g}_{kk}\}$, of the $L \times L$ Hermitian complex matrix $\mathbb{E}\{\mathbf{g}_{ki} \mathbf{g}_{ki}^{\mathrm{H}}\}$, and of the real-valued $L \times L$ diagonal matrix $\mathbf{D}_k$ for $k, i \in \{1, \ldots, K\}$. Hence, $KL + (L^2 \ K^2 + KL)/2$ complex scalars are needed in total. These values are summarized in Table I.

C. Level 2: Local Processing & Simple Centralized Decoding

Although the optimized LSFD step in Level 3 gives the highest SE among schemes with local combining at each AP, it requires knowledge of a number of statistical parameters that grows quadratically with L and K, which can be very large in Cell-free mMIMO. In practice, this large number of parameters need to be jointly estimated by the APs and sent to the CPU. This might not be feasible, especially if the statistics vary with time. To overcome this issue, the CPU can alternatively create its estimate of the signal s_k from UE k by simply taking the average of the local estimates, as proposed

[5]When dealing with ergodic capacities, all deterministic parameters can be assumed known without loss of generality, because these can be estimated using a finite number of transmission resources, while the capacity is only achieved as the amount of transmission resources goes to infinity. Hence, the estimation overhead for obtaining deterministic parameters is negligible.

in the early papers on the topic [16], [17].[6] This yields

$$\hat{s}_k = \frac{1}{L}\sum_{l=1}^{L} \check{s}_{kl} \tag{23}$$

where $\check{s}_{kl}$ is given in (15) and can be obtained by any local combining vector. Since this is equivalent to setting $\mathbf{a}_k = [1/L, \ldots, 1/L]^{\mathrm{T}}$ in Proposition 2, the following result is obtained.

Corollary 3*: At Level 2, an achievable SE of UE k is*

$$\mathsf{SE}_k^{(2)} = \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2\left(1 + \mathsf{SINR}_k^{(2)}\right) \tag{24}$$

with the effective SINR given in (25) on the top of next page, where the expectations are taken with respect to all sources of randomness.

As for Proposition 2, the above SE can be utilized along with any local combining vector and also channel estimator. If MR is used with single-antenna APs (i.e., $N = 1$), then Corollary 3 reduces to the case considered in [16] and can be computed in closed form (similar results are found in [31], [34]–[38]). The number of complex scalars to be exchanged per coherence block is the same as at Level 3. The key difference is that no statistical parameters are needed at the CPU. This is summarized in Table I.

D. Level 1: Small-Cell Network

The simplest implementation level is when the signal from UE k is decoded by using only the received signal from one AP. In this case, the decoding can be done locally at the AP by using its own local channel estimates without exchange anything with the CPU.[7] This makes the network truly distributed [15, Sec. 4.2] and essentially turns Cell-free mMIMO into a small-cell network. The macro diversity achieved by selecting the best out of many APs could potentially make it competitive compared to conventional Cellular mMIMO with larger cells.

Cell-free mMIMO and small cells were compared in [16], [17] with $N = 1$ and an AP selection based on the largest large-scale fading coefficient β_{kl}. In addition to this, the authors impose that each AP can only serve one UE. Unlike [16], [17], we remove all these restrictions by assuming an arbitrary number of antennas per AP and letting the AP that gives the highest SE to a specific UE be responsible for decoding its signal. The latter makes the AP association more complex than [16], [17], but the numerical results in Section IV show that it vastly improves the performance.[8] Within the above setting, the following result is obtained.

Corollary 4*: At Level 1, an achievable SE of UE k is*

$$\mathsf{SE}_k^{(1)} = \left(1 - \frac{\tau_p}{\tau_c}\right) \max_{l \in \{1,\ldots,L\}} \mathbb{E}\left\{\log_2\left(1 + \mathsf{SINR}_{kl}^{(1)}\right)\right\} \tag{26}$$

where the instantaneous effective SINR at AP l is

$$\mathsf{SINR}_{kl}^{(1)} = \frac{p_k |\mathbf{v}_{kl}^{\mathrm{H}} \hat{\mathbf{h}}_{kl}|^2}{\sum\limits_{\substack{i=1 \\ i \neq k}}^{K} p_i |\mathbf{v}_{kl}^{\mathrm{H}} \hat{\mathbf{h}}_{il}|^2 + \mathbf{v}_{kl}^{\mathrm{H}} \left(\sum\limits_{i=1}^{K} p_i \mathbf{C}_{il} + \sigma^2 \mathbf{I}_N\right) \mathbf{v}_{kl}} \tag{27}$$

and the expectation is with respect to the channel estimates. The maximum value in (27) *is achieved with the L-MMSE combining in* (16) *and is given by*

$$\mathsf{SINR}_{kl}^{(1)} = p_k \hat{\mathbf{h}}_{kl}^{\mathrm{H}} \left(\sum_{\substack{i=1 \\ i \neq k}}^{K} p_i \hat{\mathbf{h}}_{il} \hat{\mathbf{h}}_{il}^{\mathrm{H}} + \sum_{i=1}^{K} p_i \mathbf{C}_{il} + \sigma^2 \mathbf{I}_N\right)^{-1} \hat{\mathbf{h}}_{kl}. \tag{28}$$

Proof: For each AP, the SE is computed in the same way as in Proposition 1 and the maximum SINR is achieved as in Corollary 1. ∎

The SE expression above is more general than the one considered for small cells in [16], where $N = 1$ is considered and each AP only estimates the channel of the UE it serves. When considering that special case, the following result is obtained instead.

Proposition 3*: At Level 1 with $N = 1$, if AP l decodes the signal from UE k using only its local estimate $\hat{\mathbf{h}}_{kl}$, an achievable SE is*

$$\frac{e^{\frac{1}{\omega_{kl}(1+A_{kl})}} E_1\left(\frac{1}{\omega_{kl}(1+A_{kl})}\right) - e^{\frac{1}{\omega_{kl}A_{kl}}} E_1\left(\frac{1}{\omega_{kl}A_{kl}}\right)}{\ln(2)} \tag{29}$$

where $A_{kl} = \sum\limits_{i \in \mathcal{P}_k \setminus \{k\}} \left(\frac{p_i \beta_{il}}{p_k \beta_{kl}}\right)^2$ is due to pilot contamination,

$$\omega_{kl} = \frac{p_k^2 \tau_p \beta_{kl}^2}{\Psi_{t_k l}\left(\sum\limits_{i \notin \mathcal{P}_k} p_i \beta_{il} + \sum\limits_{i \in \mathcal{P}_k} p_i C_{il} + \sigma^2\right)}, \tag{30}$$

$E_1(x) = \int_1^\infty \frac{e^{-xu}}{u} du$ denotes the exponential integral, and $\ln(\cdot)$ denotes the natural logarithm.

Proof: The proof is given in Appendix B. ∎

Comparing the achievable SE in (29) with [16, Eq. (47)] we notice that, despite the different notation, the equivalence only holds when $A_{kl} = 0$; that is, in the absence of pilot contamination. Although [16] states the result without proof, it seems that the paper has neglected the conditioning on the local channel estimate $\hat{\mathbf{h}}_{kl}$ when computing the interference power; see (47) in Appendix B. This leads to an approximate SE rather than an exact expression. This is why we included Proposition 3 in this paper and will use it for numerical comparison in Section IV.

Remark 3*: We noticed that the expression in* (29) *is numerically unstable when $\omega_{kl}(1 + A_{kl})$ and/or $\omega_{kl}A_{kl}$ are small. This is because $e^{1/x} \to \infty$ and $E_1(1/x) \to 0$ when $x \to 0$. When this happens, one can instead utilize the bounds*

[6] Level 2 also includes other cases where the weight a_{kl} is selected based only on the statistical information available at AP l. For example, we have tried $a_{kl} = \beta_{lk}^{\nu}$ for different exponents ν but the performance gap to Level 3 remained to be large. Further research in this direction is needed.

[7] In all the four levels, the K data streams need to be transmitted to the core network after decoding. This requires a backhaul load proportional to the sum SE, which is not included in Table I but is different for each level.

[8] In practice, selecting the AP that maximizes the SE can be replaced by selecting the AP that maximizes some kind of approximate closed-form SINR. Such a selection rule has the same implementation complexity as selecting the AP with the largest large-scale fading coefficient. However, the challenge is that the SINR is affected by the transmit powers, so if these powers are optimized, the optimization must also involve the AP selection.

$$\mathsf{SINR}_k^{(2)} = \frac{p_k \left| \sum_{l=1}^{L} \mathbb{E}\left\{ \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{h}_{kl} \right\} \right|^2}{\sum_{i=1}^{K} p_i \mathbb{E}\left\{ \left| \sum_{l=1}^{L} \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{h}_{il} \right|^2 \right\} - p_k \left| \sum_{l=1}^{L} \mathbb{E}\{ \mathbf{v}_{kl}^{\mathrm{H}} \mathbf{h}_{kl} \} \right|^2 + \sigma^2 \sum_{l=1}^{L} \mathbb{E}\{ \|\mathbf{v}_{kl}\|^2 \}} \tag{25}$$

$\frac{x}{1+x} \leq e^{1/x} E_1(1/x) \leq x$ *in [49, Eq. (5.1.19)] to realize that* $e^{1/x} E_1(1/x) \approx x$ *when* $x \to 0$.

IV. Cell-Free Versus Cellular mMIMO

In this section, we compare the uplink performance of Cell-free mMIMO, with the different cooperation levels and either MR or MMSE/L-MMSE combining, and Cellular mMIMO. We first briefly describe the cellular setup that is considered.

A. Cellular mMIMO Setup

We consider a cellular network with $L_{\mathrm{c}} = 4$ cells, $M_{\mathrm{c}} = 100$ antennas per cellular BS, and $K_{\mathrm{c}} = 10$ UEs per cell. The block-fading channel from BS j to UE k in cell l is modeled as

$$\mathbf{h}_{lk}^{j} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \mathbf{R}_{lk}^{j} \right) \tag{31}$$

where $\mathbf{R}_{lk}^{j} \in \mathbb{C}^{M_{\mathrm{c}} \times M_{\mathrm{c}}}$ is the spatial correlation matrix with large-scale fading coefficient $\beta_{lk}^{j} \triangleq \mathrm{tr}(\mathbf{R}_{lk}^{j})/M_{\mathrm{c}}$ describing the geometric pathloss and shadowing. The uplink transmit power of UE k in cell l is denoted by $p_{lk} \geq 0$.

We assume there are $\tau_p = K_{\mathrm{c}}$ mutually orthogonal pilots and that UE k in every cell uses the same pilot (i.e., pilot reuse one). When using standard MMSE estimation [3, Th. 3.1], the MMSE estimate of $\mathbf{h}_{lk}^{j} \in \mathbb{C}^{M_{\mathrm{c}}}$ is given by

$$\hat{\mathbf{h}}_{li}^{j} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \mathbf{R}_{li}^{j} - \mathbf{C}_{li}^{j} \right) \tag{32}$$

and the independent estimation error $\tilde{\mathbf{h}}_{li}^{j} \in \mathbb{C}^{M_{\mathrm{c}}}$ is

$$\tilde{\mathbf{h}}_{li}^{j} \triangleq \mathbf{h}_{li}^{j} - \hat{\mathbf{h}}_{li}^{j} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \mathbf{C}_{li}^{j} \right) \tag{33}$$

with

$$\mathbf{C}_{li}^{j} = \mathbf{R}_{li}^{j} - p_{li} \tau_p \mathbf{R}_{li}^{j} \left(\sum_{l'=1}^{L_{\mathrm{c}}} p_{l'i} \tau_p \mathbf{R}_{l'i}^{j} + \sigma^2 \mathbf{I}_{M_{\mathrm{c}}} \right)^{-1} \mathbf{R}_{li}^{j}. \tag{34}$$

An achievable SE of UE k in cell j is [3, Th. 4.1]

$$\mathsf{SE}_{jk}^{(\mathrm{c})} = \left(1 - \frac{\tau_p}{\tau_c} \right) \mathbb{E} \left\{ \log_2 \left(1 + \mathsf{SINR}_{jk}^{(\mathrm{c})} \right) \right\} \tag{35}$$

where the effective SINR, $\mathsf{SINR}_{jk}^{(\mathrm{c})}$, is maximized by multi-cell MMSE (M-MMSE) combining [9]. This gives

$$\begin{aligned} &\mathsf{SINR}_{jk}^{(\mathrm{c})} \\ &= p_{jk} (\hat{\mathbf{h}}_{jk}^{j})^{\mathrm{H}} \left(\sum_{l=1}^{L_{\mathrm{c}}} \sum_{\substack{i=1 \\ (l,i) \neq (j,k)}}^{K_{\mathrm{c}}} p_{li} \hat{\mathbf{h}}_{li}^{j} (\hat{\mathbf{h}}_{li}^{j})^{\mathrm{H}} \right. \\ &\qquad \left. + \sum_{l=1}^{L_{\mathrm{c}}} \sum_{i=1}^{K_{\mathrm{c}}} p_{li} \mathbf{C}_{li}^{j} + \sigma^2 \mathbf{I}_{M_{\mathrm{c}}} \right)^{-1} \hat{\mathbf{h}}_{jk}^{j}. \end{aligned} \tag{36}$$

Other combining schemes can be used but they provide lower SEs. By considering M-MMSE, we thus compare Cell-free mMIMO with the most competitive form of Cellular mMIMO.

B. Simulation Setup and Propagation Model

The cellular network has 4 square cells in a $1 \times 1\,\mathrm{km}$ area, as in Fig. 1, with 100 co-located antennas per BS. The cell-free network is deployed in the same area and has either 400 single-antenna APs (i.e., $N = 1$) or 100 four-antenna APs (i.e., $N = 4$). Hence, all the network configurations have the same number of antennas. To make a fair comparison, the APs are deployed on a square grid (we consider random deployment later in this section) and the same propagation model is used in all cases. We anticipate that the APs in Cell-free mMIMO will be deployed in urban environments with high user loads, roughly 10 m above the ground. This matches well with the 3GPP Urban Microcell model in [50, Table B.1.2.1-1] with a 2 GHz carrier frequency and

$$\beta_{kl}\,[\mathrm{dB}] = -30.5 - 36.7 \log_{10} \left(\frac{d_{kl}}{1\,\mathrm{m}} \right) + F_{kl} \tag{37}$$

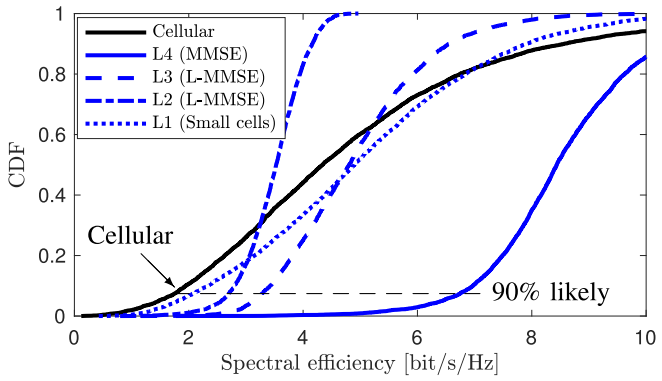
where d_{kl} is the distance between UE k and AP l (computed as the minimum over different wrap-around cases, and taking the 10 m height difference into account) and $F_{kl} \sim \mathcal{N}(0, 4^2)$ is the shadow fading. The shadowing terms from an AP to different UEs are correlated as [50, Table B.1.2.2.1-4]

$$\mathbb{E}\{F_{kl} F_{ij}\} = \begin{cases} 4^2 2^{-\delta_{ki}/9\,\mathrm{m}}, & l = j \\ 0, & l \neq j \end{cases} \tag{38}$$

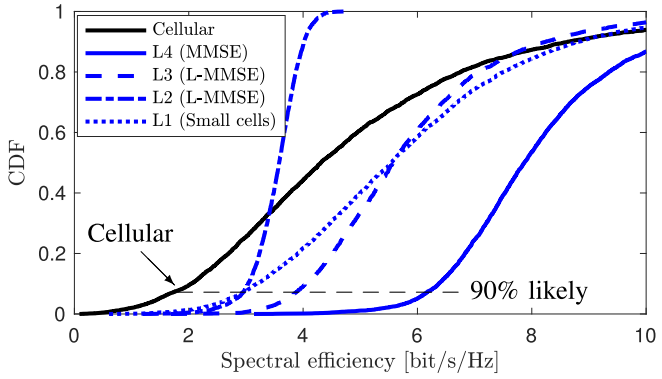
where δ_{ki} is the distance between UE k and UE i. The second row in (38) accounts for the correlation of shadowing terms related to two different APs, which is negligible since we have at least 50 m between adjacent APs in the simulation setup (notice that $2^{-50/9} \approx 0.02$).

Since the propagation model from [50] is designed for cellular networks, we use the same propagation model for Cellular mMIMO by simply adding an additional index to all the parameters to specify in which cell a particular UE resides. By having a common model for cell-free and cellular networks, we can be sure that the performance differences that we observe are caused by differences in technology characteristics, and not by the propagation model. There are $K = 40$ UEs in the simulation setup, whereof ten are uniformly dropped in each cell and assigned to unique pilots with random indices.[9] The same UE locations and pilot assignments are used in the cell-free case, but the shadowing is generated independently.

[9]Each UE in the Cellular mMIMO case is connecting to the BS providing the largest large-scale fading coefficient; that is, $\beta_{jk}^{j} = \max_l \beta_{jk}^{j}$. Due to the shadowing, this might not be the geographically closest BS.



(a) Cell-free with $L = 400$, $N = 1$.



(b) Cell-free with $L = 100$, $N = 4$.

Fig. 2. Comparison of cellular mMIMO and cell-free mMIMO when using MMSE or L-MMSE combining.

The cellular BSs and multi-antenna APs are equipped with half-wavelength-spaced uniform linear arrays. The spatial correlation is generated using the Gaussian local scattering model with 15° angular standard deviation [3, Sec. 2.6]. All UEs transmit with power $p_k = p_{jk} = 100\,\text{mW}$, the bandwidth is 20 MHz, the noise power is $\sigma^2 = -96\,\text{dBm}$, and the coherence blocks contain $\tau_c = 200$ channel uses (e.g., achieved by 2 ms coherence time and 100 kHz coherence bandwidth).

Remark 4*: The early Cell-free mMIMO papers [16], [24] used another propagation model, which has since then become standard in the field. However, that model is based on the COST-Hata model from [51] for macro-cells, where the APs are at least 30 m above the ground and the UEs are at least 1 km from the AP. This is very different from the micro-cell-like deployment we anticipate for Cell-free mMIMO and it should be noted that the model creators themselves specified that it "must not be used for micro-cells" [51, Ch. 4]. Moreover, the model in [16], [24] has no shadowing when a UE is closer than 50 m from an AP, which is often the case in Cell-free mMIMO deployments. When the distance is larger, the shadowing decorrelation distance is* $10\times$ *larger than in the 3GPP model [50]. For all these reasons, we believe that the propagation model used in this paper is a better baseline for evaluating Cell-free mMIMO systems.*

C. Numerical Comparisons

Fig. 2(a) compares Cellular mMIMO and Cell-free mMIMO with $L = 400$ and $N = 1$. The figure shows the cumulative distribution function (CDF) of the SE of a randomly located UE, when using MMSE or L-MMSE combining in the cell-free cases. At the 90% or 95% likely SE points (i.e., where the vertical axis is 0.1 or 0.05), the cell-free cases perform according to their level: Level 4 provides by far the highest SEs, while Level 1 gives the lowest SEs but is anyway preferable as compared to Cellular mMIMO. Looking at the complete CDF curves, the situation is more complicated since the Level 1 and Cellular mMIMO curves are crossing the Level 2 and Level 3 curves. Hence, UEs with good channel conditions get better performance with these methods. However, Level 4 performs better than Cellular mMIMO for every UE.

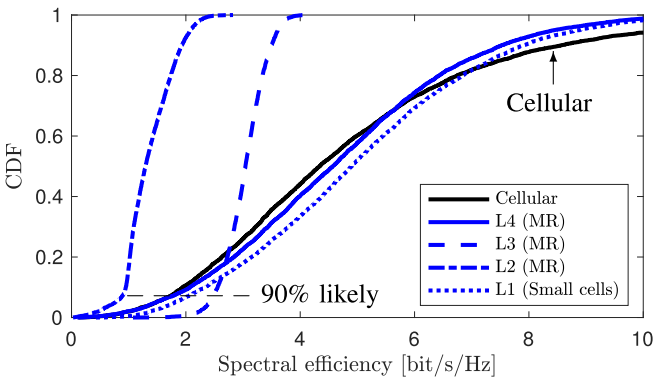


Fig. 3. Comparison of cellular mMIMO with cell-free ($L = 400$, $N = 1$) when using MR combining.

Fig. 2(b) considers the same setup but with fewer APs that are equipped with multiple antennas: $L = 100$ and $N = 4$. The general trends are the same as in Fig. 2(a) but Level 4 loses in SE due to the reduced macro diversity; that is, the average distance from a UE to an AP is increased. In contrast, Level 1 gains in performance since each AP can now suppress interference locally, by using its four antennas. In fact, Level 1 is now comparable to Level 2 for the weakest UEs and substantially better for the strongest UEs.

Next, Fig. 3 considers the case $L = 400$, $N = 1$ and MR combining, which is the receiver processing advocated in the early papers on Cell-free mMIMO. More precisely, Level 2 was considered in [16] and Level 3 in [24]. The poor processing leads to a large SE loss, compared to Fig. 2(a), for all levels of Cell-free mMIMO receiver cooperation, except Level 1. In fact, Level 2 is outperformed by both small cells (Level 1) and Cellular mMIMO for every single UE. Note that we are considering single-antenna APs in this figure, so MR processing is suboptimal even in that basic case, and the use of LSFD in Level 3 cannot make up for the performance loss. This is because L-MMSE and MR differ by a non-deterministic scalar and LSFD only involves deterministic scalars. Not even Level 4 performs better than Cellular mMIMO or small cells when using MR, so we can conclude that Cell-free mMIMO should never use the MR scheme.

D. Revisiting "Cell-Free Massive MIMO Versus Small Cells"

Interestingly, our observations in Fig. 3 contradict the previous results in [16], where Cell-free mMIMO with

'Level 2 (MR)' was shown to perform much better than small cells, in terms of both 95%-likely and median SE. The reason for the differences is explained in this subsection by reproducing [16, Fig. 4, Fig. 6] and adding some additional curves to them. The following three-slope propagation model was used in [16]:

$$\beta_{kl}\,[\text{dB}] = \begin{cases} -81.2, & d_{kl} < 10\,\text{m} \\ -61.2 - 20\log_{10}\left(\frac{d_{kl}}{1\,\text{m}}\right), & 10\,\text{m} \leq d_{kl} < 50\,\text{m} \\ -35.7 - 35\log_{10}\left(\frac{d_{kl}}{1\,\text{m}}\right) + F_{kl}, & d_{kl} \geq 50\,\text{m} \end{cases} \tag{39}$$

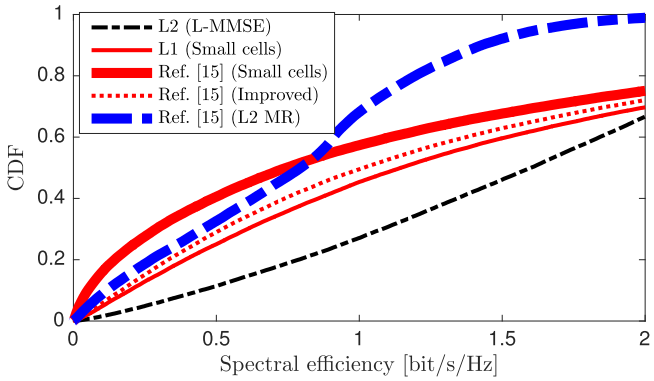
where d_{kl} is the horizontal distance between UE k and AP l (i.e., ignoring the height difference). The shadowing term $F_{kl} \sim \mathcal{N}(0, 8^2)$ only appears when the distance is larger than 50 m and the terms are correlated as

$$\mathbb{E}\{F_{kl}F_{ij}\} = \frac{8^2}{2}\left(2^{-\delta_{ki}/100\,\text{m}} + 2^{-\varrho_{lj}/100\,\text{m}}\right) \tag{40}$$

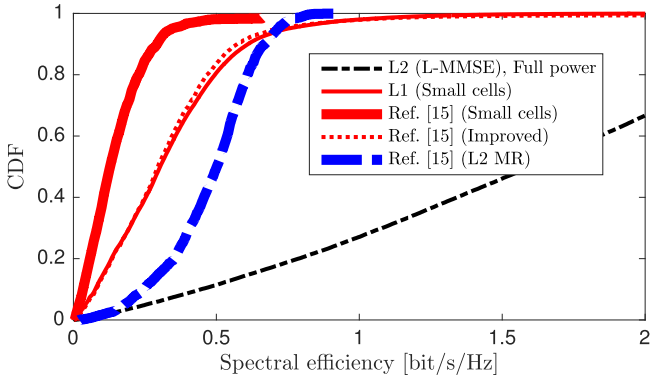
where δ_{ki} is the same as in (38) and ϱ_{lj} is the distance between AP l and AP j. The maximum UE power is 100 mW, the bandwidth is 20 MHz, the noise power is $\sigma^2 = -92$ dBm, and the coherence blocks are determined by $\tau_c = 200$.

We consider the same setup as in [16] with $L = 100$ uniformly distributed APs in a 1×1 km area, $N = 1$ antenna per AP, $K = 40$ uniformly distributed UEs, and $\tau_p = 20$ orthogonal pilots. The pilots are assigned to the UEs according to the greedy algorithm described in [16, Sec. IV.A], with the only difference that we use the uplink SE as the metric in Step 2 of the algorithm (instead of the downlink SE). Moreover, we use Proposition 3 to accurately compute the SE with small cells, but this has little impact on the results. The thick lines in Fig. 4 correspond to the original curves from [16] with correlated shadowing. Fig. 4(a) considers the case when the UEs transmit at full power (as in [16, Fig. 6]) and Fig. 4(b) considers the case when the UEs transmit pilots at full power but reduce the power during the data transmission to optimize the network-wide max-min fairness (as in [16, Fig. 4]). To this end, we use the same optimization algorithms as in [16].

In the full power case, in Fig. 4(a), it is clear from the thick curves that Cell-free mMIMO at Level 2 with MR gives the UEs with the 50% worst channel conditions substantially higher SE than with small cells. The remaining UEs get better SEs with small cells, which indicates that the considered Cell-free system is not well implemented—since the cell-free network has access to more APs and signal observations when decoding a UE's signal, the performance should be better for *everyone*. Moreover, the comparison in [16, Fig. 6] is based on a suboptimal assignment of UEs to small cells; the UEs are sequentially selecting the AP that has the largest large-scale fading coefficient β_{kl} (i.e., the best channel), but only among those that are not already serving another UE. If we change that to let each UE being served by the AP giving the highest SE, represented by the curve 'Ref. [14] (Improved)', then the performance gap between Cell-free mMIMO and small cells diminishes. The reason is that around 40% of the UEs prefer to be served by another small cell. Additionally, if we use the new improved SE expression in Corollary 4, represented by the curve 'L1 (Small cells)', all the UEs get higher SE with small cells than with Cell-free mMIMO.



(a) Full power.



(b) Max-min power control.

Fig. 4. Comparison of cell-free mMIMO at Level 2 and small cells, using different SE expressions and AP assignments. The UEs either transmit with full power or optimizes the power as described in [16]. The thick lines correspond to the curves in [16, Fig. 4, Fig. 6].

Does this mean that small cells are actually better than Cell-free mMIMO? The answer is no. Indeed, as observed in the last subsection, the problem is that MR combining performs badly in Cell-free mMIMO, even if single-antenna APs are used. By simply changing to Level 2 with L-MMSE combining, the rightmost curve in Fig. 4(a) is achieved, which gives uniformly higher SE to all the UEs than when using small cells. Even higher SE can be achieved by considering Level 3 or Level 4 implementations.

The results in Fig. 4(b) with max-min power control are different and more in line with the observations made in [16]: Level 2 with MR gives much higher SE than small cells, but the gap can be reduced by selecting APs based on the maximum SE rather than the maximum β_{kl} (represented by the curves 'Ref. [15] (Improved)' and 'L1 (Small cells)'). The benefit of max-min power control can be seen by considering the two thick lines (obtained as in [16]): the lower end of the CDF curves are shifted to the right as compared to the full power case in Fig. 4(a), yielding a higher guaranteed SE level. Nevertheless, the use of L-MMSE combining is more appealing than the use of max-min power control, as can be seen

from the rightmost curve in Fig. 4(b) that considers Level 2 with L-MMSE and full power transmission. This approach gives the same performance as 'L2 MR' with max-min fairness for the 2% weakest UEs, but higher SE for all other UEs; for example, it achieves a 40% higher 95%-likely SE and a $3\times$ higher median SE. Hence, if L-MMSE processing is used, advanced power control is not needed in Cell-free mMIMO to give good performance to the weakest UEs.

V. Level 4 With Non-Linear Decoding

Section IV-C showed that Level 4 can provide vastly higher SE than the other cooperation levels in Cell-free mMIMO. The comparison is based on using linear receive combining, but another benefit of centralizing the signal processing at a CPU is that more advanced decoding methods can potentially be used, since network-wide CSI and high computational resources are available. In this section, we investigate the potential benefits of the non-linear successive interference cancelation (SIC) method [52, Sec. 8.3.4] in Cell-free mMIMO, which means that the CPU decodes one UE signal at a time, and then sequentially subtracts interference that the decoded signal caused to the remaining signals. The interference cannot be fully canceled since the CPU has imperfect CSI, but it can still improve the SE of the UEs compared to linear combining.

Proposition 4*: At Level 4, if the MMSE estimator is used to compute channel estimates for all UEs and the signals are decoded using MMSE combining and SIC (MMSE-SIC), then for any decoding order an achievable sum SE is*

$$\mathsf{SSE}^{(\mathrm{SIC})} = \left(1 - \frac{\tau_p}{\tau_c}\right) \mathbb{E}\left\{\log_2 \det\left(\mathbf{I}_K + \mathbf{P}\hat{\mathbf{H}}^{\mathrm{H}}\mathbf{E}^{-1}\hat{\mathbf{H}}\right)\right\} \tag{41}$$

where $\mathbf{P} = \mathrm{diag}(p_1, \ldots, p_K)$, $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1 \ \ldots \ \hat{\mathbf{h}}_K] \in \mathbb{C}^{LN \times K}$, $\mathbf{E} = \sum_{i=1}^{K} p_i \mathbf{C}_i + \sigma^2 \mathbf{I}_{LN}$, *and the expectation is with respect to the channel estimates.*

Proof: The proof is given in Appendix C. ∎

Proposition 4 provides the sum SE of the Cell-free mMIMO network, and not the individual SEs of the UEs. The reason is that the latter depends on the decoding order; that is, the later a UE is decoded, the less interference it will be affected by and thereby it will gain more in SE compared to using linear combining. Irrespective of the decoding order, all UEs get at least as high SE with MMSE-SIC as with MMSE combining.

Fig. 5 revisits the scenario in Fig. 2(a) by considering Cell-free mMIMO with $L = 400$ and $N = 1$. The CDF of the sum SE over different random realizations of the UE locations is plotted when using either Levels 1-4 with MMSE/L-MMSE combining or Level 4 with MMSE-SIC, based on Proposition 4. The MMSE-SIC method improves the sum SE, but the average gain over 'L4 (MMSE)' is only 1%. The reason for such modest gain is the favorable propagation phenomenon that makes the UEs' channels nearly orthogonal [27], meaning that the inter-user interference is effectively canceled by the MMSE processing described in Section III. Hence, we conclude that non-linear processing is not needed in Cell-free mMIMO. This is also the reason why we did not present the detailed per-user SEs in this section.

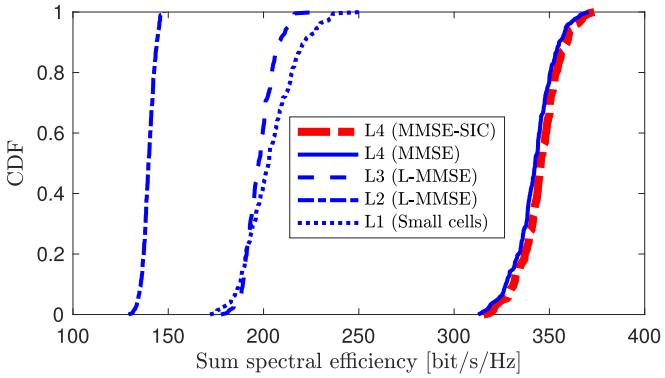


Fig. 5. CDF of the sum SE over different random user locations with $L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$. The four cooperation levels are compared with MMSE-SIC, based on Proposition 4.

Another observation that can be made from Fig. 5 is that Level 1 and Level 3 provide roughly the same sum SE, while Level 2 is far behind in performance. The large gap to Level 4 further reinforces the point that a centralized implementation is strongly preferred in Cell-free mMIMO.

VI. A Look at the Fronthaul Signaling Load

The reported results show that a Level 4 implementation is strongly preferred. The counterargument might be that such an implementation would require much more fronthaul signaling than Level 2 and Level 3, but we will now show that it is not necessarily the case. By using the formulas in Table I, Level 4 requires less signaling if

$$\frac{\tau_c NL}{(\tau_c - \tau_p)KL} = \frac{\tau_c}{\tau_c - \tau_p}\frac{N}{K} < 1. \tag{42}$$

Since $\frac{\tau_c}{\tau_c - \tau_p} \approx 1$ and $K \gg N$ are typical for Cell-free mMIMO, Level 4 actually requires *much less* signaling.

Fig. 6 shows how many complex scalars need to be sent from an AP to the CPU per channel use, as a function of the coherence block length τ_c. We consider the same setup as in Fig. 2(b): $L = 100$, $N = 4$, $K = 40$, and $\tau_p = 10$. Level 4 requires more signaling if $\tau_c \leq 11$, while much less signaling is required when τ_c becomes a hundred, as in practical systems. As $\tau_c \to \infty$, Level 2 and 3 require $K/N = 10$ times more fronthaul signaling than Level 4. The reason is that the received data signals constitute a much larger number of scalars than the channel estimates. Since $K \geq N$ is typically the case in Cell-free mMIMO, Level 2 and Level 3 increase the fronthaul signaling by processing the N-dimensional vector $\mathbf{y}_l$ into the K-dimensional vector $[\check{s}_{1l}, \ldots, \check{s}_{Kl}]^{\mathrm{T}}$. In practice, an AP will not serve all the UEs in the network but only those with a good channel. Nevertheless, as long as each AP serves more UEs than it has antennas (e.g., more than one UE in conventional Cell-free mMIMO with $N = 1$), Level 4 is preferable in terms of fronthaul signaling.

Admittedly, this comparison assumes that all scalars are shared with infinite precision, while in practice it is plausible that the pilot signals require higher bit-resolution when sent to the CPU than the data signals. On the other hand, the pilot signals constitute only a minor fraction of the total signaling and [32], [53] recently showed that the estimates can be compressed rather well.

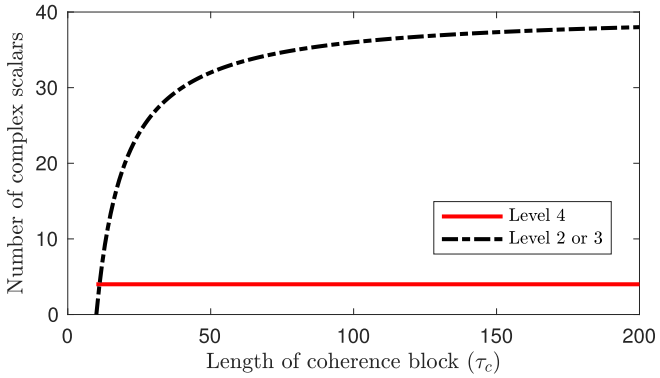


Fig. 6. Number of complex scalars that needs to be shared between an AP and the CPU per channel use ($L = 100$, $N = 4$, $K = 40$, $\tau_p = 10$).

A. Serial Fronthaul

Since Level 2 and Level 3 necessarily provide lower SE than Level 4, these levels are only practically interesting if they require lower fronthaul capacity. The previous example shows that this is not the case when each AP transmits individually over the fronthaul, but there are alternative solutions that reduce the fronthaul capacity requirement. In particular, this happens when several APs are deployed along the same wired connection, as illustrated in the lower right corner of Fig. 1(b).

Suppose AP 1 and AP 2 share a fronthaul connection in this way. When the locally estimated signal $\check{s}_{k1}$ at AP 1 is sent over the fronthaul to AP 2, this AP will compute $\check{s}_{k1} + \check{s}_{k2}$. The result is then sent to the CPU, which can still form its signal estimate in (23) at Level 2, since it is the summation of the local estimates at all the APs. By instead transmitting the weighted local estimates $a_{kl}\check{s}_{kl}$ over the fronthaul, Level 3 can be implemented in the same sequential fashion (assuming that a_{kl} can be computed locally at AP l).

Since only one scalar per UE is transmitted over each segment of the fronthaul, the capacity requirement does not grow with the number of APs that are sharing the wired connection. In the extreme case when all APs are deployed along the same wire, the number of complex scalars sent over the fronthaul per coherence block reduces from $(\tau_c - \tau_p)KL$ in Table I to $(\tau_c - \tau_p)K$. This type of serial fronthaul is needed for Level 2 and Level 3 to make practical sense, which is why it is adopted by the radio stripes concept described in [43].

VII. Conclusion

This paper introduced a taxonomy for Cell-free mMIMO with four different implementation levels, from fully centralized to fully distributed, and generalized previous results to account for multi-antenna APs, spatially correlated fading, and arbitrary receive combining. The majority of previous papers on this topic relied upon a distributed implementation with local MR processing. Remarkably, we discovered that this is basically the worst way to operate cell-free networks.

Firstly, local MMSE processing provides substantially higher SE than MR, and is the key prerequisite for Cell-free mMIMO to outperform conventional Cellular mMIMO and small-cell networks. Importantly, this is the case even if each AP is equipped with only one antenna; local MMSE processing can roughly double the SE per UE.

Secondly, we showed that a centralized implementation, with all the signal processing taking place at an edge-cloud processor (a.k.a. CPU in the cell-free literature), is highly preferable compared to distributed alternatives. In fact, the centralized Level 4 implementation can simultaneously increase the SE and reduce the fronthaul signaling. Linear processing is sufficient at Level 4 since non-linear processing provides negligible gains due to the favorable propagation property [27]. The pCell technology [19] is an example of a centralized cell-free system, which demonstrates that it is practically feasible. A serial fronthaul is needed to make a distributed implementation competitive in terms of the fronthaul capacity requirements, and an improved version of Level 3 needs to be developed to reduce the performance gap to Level 4. Non-linear processing can be useful at Level 3 and the compute-and-forward relaying framework can potential guide the development of such methods [47], [54], [55].

An interesting analogy can be made between the results in this paper and recent developments in the Cellular mMIMO area. The seminal paper [4] advocated for using MR processing, based on asymptotic arguments. MR was known to be suboptimal when having a small number of antennas, but anyway became the most well-studied method in the literature since the SE can be computed in closed-form, even with more complicated system models containing spatially correlated fading and/or hardware impairments [3]. However, recent works have shown that M-MMSE processing greatly outperforms MR even in the asymptotic regime [9]. Similarly, the main conclusion of this paper is that it is time to forget about MR also in Cell-free mMIMO and instead consider only MMSE-based schemes—irrespective of the level of cooperation among the APs and the number of antennas used at each one.

Appendix A
Proof of Proposition 2

Since the CPU does not have knowledge of the channel estimates, it needs to treat the average channel gain $\mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}$ as the true deterministic channel. Hence, the signal model is

$$\hat{s}_k = \mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}s_k + \upsilon_k \tag{43}$$

which is a "deterministic" channel with the additive interference plus noise term

$$\upsilon_k = \left(\mathbf{a}_k^{\mathrm{H}}\mathbf{g}_{kk} - \mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}\right)s_k + \sum_{i=1, i\neq k}^{K} \mathbf{a}_k^{\mathrm{H}}\mathbf{g}_{ki}\mathbf{g}_{ki}^{\mathrm{H}}\mathbf{a}_k s_i + \mathbf{n}_k'. \tag{44}$$

The interference term υ_k has zero mean and is uncorrelated with the signal term in (43) since

$$\underbrace{\mathbb{E}\{\mathbf{a}_k^{\mathrm{H}}\mathbf{g}_{kk} - \mathbf{a}_k^{\mathrm{H}}\mathbb{E}\{\mathbf{g}_{kk}\}\}}_{=0}\mathbb{E}\{|s_k|^2\} = 0. \tag{45}$$

Therefore, we can apply [3, Cor. 1.3] to obtain the achievable SE in (19).

Appendix B
Proof of Proposition 3

In this proof, we drop the bold face to emphasize that all parameters are scalars. Using the capacity lower bound

in [3, Cor. 1.3] with $\hat{h}_{kl}$ as the known channel realization, an achievable SE is

$$\mathbb{E}\left\{\log_2\left(1+\frac{p_k|\hat{h}_{kl}|^2}{\mathbb{E}\{|\upsilon|^2\,|\hat{h}_{kl}\}+\sigma^2}\right)\right\} \tag{46}$$

where $\upsilon=\tilde{h}_{kl}s_k+\sum_{i\neq k}h_{il}s_i$ and

$$\mathbb{E}\{|\upsilon|^2\,|\hat{h}_{kl}\}=\sum_{i\in\mathcal{P}_k\setminus\{k\}}\frac{p_i^2\beta_{il}^2}{p_k\beta_{kl}^2}|\hat{h}_{kl}|^2+\sum_{i\notin\mathcal{P}_k}p_i\beta_{il}+\sum_{i\in\mathcal{P}_k}p_iC_{il} \tag{47}$$

by exploiting the fact that $\hat{h}_{il}$ and $\hat{h}_{kl}$ are independent for all $i\notin\mathcal{P}_k$ and $\hat{h}_{il}=\frac{\sqrt{p_i}\beta_{il}}{\sqrt{p_k}\beta_{kl}}\hat{h}_{kl}$ for all $i\in\mathcal{P}_k$. By inserting (47) into (46), we can expand the expression as

$$\mathbb{E}\left\{\log_2\left(1+|\hat{h}_{kl}|^2\frac{p_k\,(1+A_{lk})}{\sum\limits_{i\notin\mathcal{P}_k}p_i\beta_{il}+\sum\limits_{i\in\mathcal{P}_k}p_iC_{il}+\sigma^2}\right)\right\}-\mathbb{E}\left\{\log_2\left(1+|\hat{h}_{kl}|^2\frac{p_kA_{lk}}{\sum\limits_{i\notin\mathcal{P}_k}p_i\beta_{il}+\sum\limits_{i\in\mathcal{P}_k}p_iC_{il}+\sigma^2}\right)\right\} \tag{48}$$

and compute each of the expectations using [23, Lemma 3] and $\hat{h}_{kl}\sim\mathcal{N}_{\mathbb{C}}(0,p_k\tau_p\beta_{kl}^2/\Psi_{t_kl})$ to obtain the final expression in (29).

Appendix C
Proof of Proposition 4

The received signal in (9) for Level 4 can be expressed as

$$\mathbf{y}=\sum_{i=1}^{K}\hat{\mathbf{h}}_is_i+\mathbf{e} \tag{49}$$

where $\mathbf{e}\triangleq\mathbf{n}+\sum_{i=1}^{K}\tilde{\mathbf{h}}_is_i$ has zero mean and correlation matrix $\mathbf{E}$. Since the MMSE channel estimates are known and $\mathbf{e}$ is uncorrelated with $\hat{\mathbf{h}}_is_i$ for all i, (49) can be treated as a multiple access channel with colored noise. In the worst case, in terms of mutual information, the colored noise is independent of the desired signals and Gaussian distributed. Hence, we can apply pre-whitening followed by standard results on MMSE-SIC receivers to obtain the achievable sum SE [52, Sec. 8.3.4] $\mathbb{E}\{\log_2\det(\mathbf{I}_{NM}+\mathbf{A}^{-1/2}\hat{\mathbf{H}}\mathbf{P}\hat{\mathbf{H}}^{\mathrm{H}}\mathbf{A}^{-1/2})\}$. This expression reduces to (41) by utilizing the fact that $\det(\mathbf{I}+\mathbf{BC})=\det(\mathbf{I}+\mathbf{CB})$ for any matrices $\mathbf{B},\mathbf{C}$ of compatible sizes, and by including the pre-log factor $1-\tau_p/\tau_c$ that is the fraction of channel uses used for data. Note that the sum SE expression is independent of the decoding order.

References

[1] E. Björnson and L. Sanguinetti, "Cell-free versus cellular massive MIMO: What processing is needed for cell-free to win?" in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[2] V. H. M. Donald, "Advanced mobile phone service: The cellular concept," *Bell Syst. Tech. J.*, vol. 58, no. 1, pp. 15–41, Jan. 1979.

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

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

[5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

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

[7] S. Parkvall, E. Dahlman, A. Furuskär, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Standard Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[8] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[9] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.

[10] L. Sanguinetti, E. Björnson, and J. Hoydis, "Towards massive MIMO 2.0: Understanding spatial correlation, interference suppression, and pilot contamination," 2019, *arXiv:1904.03406*. [Online]. Available: https://arxiv.org/abs/1904.03406

[11] K. T. Truong and R. W. Heath, Jr., "The viability of distributed antennas for massive MIMO systems," in *Proc. 47th Asilomar Conf. Signals, Syst. Comput.*, Monterey, CA, USA, Nov. 2013, pp. 1318–1323.

[12] E. Björnson, M. Matthaiou, and M. Debbah, "Massive MIMO with non-ideal arbitrary arrays: Hardware scaling laws and circuit-aware design," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4353–4368, Aug. 2015.

[13] W. Choi and J. G. Andrews, "Downlink performance and capacity of distributed antenna systems in a multicell environment," *IEEE Trans. Wireless Commun.*, vol. 6, no. 1, pp. 69–73, Jan. 2007.

[14] R. Irmer *et al.*, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[15] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," in *Foundations and Trends in Communications and Information Theory*, vol. 9, nos. 2–3. 2013, pp. 113–381.

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

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

[18] A. Burr, M. Bashar, and D. Maryopi, "Ultra-dense radio access networks for smart cities: Cloud-RAN, fog-RAN and "cell-free," massive MIMO," in *Proc. PIMRC*, Sep. 2018, pp. 1–5.

[19] S. Perlman and A. Forenza. (Feb. 2015). *An Introduction to PCell*. [Online]. Available: http://www.rearden.com/artemis/An-Introduction-to-pCell-White-Paper-150224.pdf

[20] S. Shamai (Shitz) and B. M. Zaidel, "Enhancing the cellular downlink capacity via co-processing at the transmitting end," in *Proc. IEEE VTS 53rd Veh. Technol. Conf.*, vol. 3, May 2001, pp. 1745–1749.

[21] S. Zhou, M. Zhao, X. Xu, J. Wang, and Y. Yao, "Distributed wireless communication system: A new architecture for future public wireless access," *IEEE Commun. Mag.*, vol. 41, no. 3, pp. 108–113, Mar. 2003.

[22] S. Venkatesan, A. Lozano, and R. Valenzuela, "Network MIMO: Overcoming intercell interference in indoor wireless systems," in *Proc. Conf. Rec. 31st Asilomar Conf. Signals, Syst. Comput.*, Nov. 2007, pp. 83–87.

[23] E. Björnson, R. Zakhour, D. Gesbert, and B. Ottersten, "Cooperative multicell precoding: Rate region characterization and distributed strategies with instantaneous and statistical CSI," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4298–4310, Aug. 2010.

[24] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and B. D. Rao, "Performance of cell-free massive MIMO systems with MMSE and LSFD receivers," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 203–207.

[25] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.

[26] H. Yang and T. L. Marzetta, "Energy efficiency of massive MIMO: Cell-free vs. cellular," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.

[27] Z. Chen and E. Björnson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, Nov. 2018.
[28] A. Ashikhmin and T. Marzetta, "Pilot contamination precoding in multi-cell large scale antenna systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2012, pp. 1137–1141.
[29] A. Adhikary, A. Ashikhmin, and T. L. Marzetta, "Uplink interference reduction in large-scale antenna systems," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2194–2206, May 2017.
[30] F. Riera-Palou, G. Femenias, A. G. Armada, and A. Pérez-Neira, "Clustered cell-free massive MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.
[31] H. Yang and E. G. Larsson, "Can massive MIMO support uplink intensive applications?" in *Proc. IEEE WCNC*, Feb. 2019, pp. 1–6.
[32] M. Bashar, H. Q. Ngo, A. G. Burr, D. Maryopi, K. Cumanan, and E. G. Larsson, "On the performance of backhaul constrained cell-free massive MIMO with linear receivers," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 624–628.
[33] H. Q. Ngo, H. Tataria, M. Matthaiou, S. Jin, and E. G. Larsson, "On the performance of cell-free massive MIMO in Ricean fading," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 980–984.
[34] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.
[35] J. Zhang, Y. Wei, E. Björnson, Y. Han, and S. Jin, "Performance analysis and power control of cell-free massive MIMO systems with hardware impairments," *IEEE Access*, vol. 6, pp. 55302–55314, 2018.
[36] Ö. Özdogan, E. Björnson, and J. Zhang, "Cell-free massive MIMO with Rician fading: Estimation schemes and spectral efficiency," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2018, pp. 975–979.
[37] W. Fan, J. Zhang, E. Björnson, S. Chen, and Z. Zhong, "Performance analysis of cell-free massive MIMO over spatially correlated fading channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.
[38] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "On the uplink max-min SINR of cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2021–2036, Apr. 2019.
[39] A. Sanderovich, O. Somekh, and S. Shamai (Shitz), "Uplink macro diversity with limited backhaul capacity," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2007, pp. 11–15.
[40] P. Marsch and G. Fettweis, "Uplink CoMP under a constrained backhaul and imperfect channel knowledge," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1730–1742, Jun. 2011.
[41] O. Simeone *et al.*, "Cooperative wireless cellular systems: An information-theoretic view," in *Foundations and Trends in Communications and Information Theory*, vol. 8, nos. 1–2. 2012, pp. 1–177.
[42] R. Rogalin *et al.*, "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1815–1831, Apr. 2014.
[43] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, p. 197, Dec. 2019.
[44] M. Morelli, C. C. J. Kuo, and M. O. Pun, "Synchronization techniques for orthogonal frequency division multiple access (OFDMA): A tutorial review," *Proc. IEEE*, vol. 95, no. 7, pp. 1394–1427, Jul. 2007.
[45] I. E. Aguerri, A. Zaidi, G. Caire, and S. Shamai (Shitz), "On the capacity of cloud radio access networks with oblivious relaying," *IEEE Trans. Inf. Theory.*, vol. 65, no. 7, pp. 4575–4596, Jul. 2019.
[46] E. Biglieri, J. Proakis, and S. Shamai (Shitz), "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619–2691, Oct. 1998.
[47] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.
[48] T. Van Chien, C. Mollén, and E. Björnson, "Large-scale-fading decoding in cellular massive MIMO systems with spatially correlated channels," *IEEE Trans. Commun.*, vol. 67, no. 4, pp. 2746–2762, Apr. 2019.
[49] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York, NY, USA: Dover, 1965.
[50] *Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, document TS 36.814, 3GPP, Mar. 2017.
[51] E. Damosso and L. M. Correia. (1999). *COST Action 231: Digital Mobile Radio Towards Future Generation Systems: Final Report*. [Online]. Available: http://www.lx.it.pt/cost231/final_report.htm
[52] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[53] D. Maryopi and A. G. Burr, "Few-bit CSI acquisition for centralized cell-free massive MIMO with spatial correlation," in *Proc. WCNC*, Feb. 2019, pp. 1–14.
[54] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.
[55] Y. Zhou, Y. Xu, W. Yu, and J. Chen, "On the optimal fronthaul compression and decoding strategies for uplink cloud radio access networks," *IEEE Trans. Inf. Theory.*, vol. 62, no. 12, pp. 7402–7418, Dec. 2016.

Emil Björnson (S'07–M'12–SM'17) received the M.S. degree in engineering mathematics from Lund University, Sweden, in 2007, and the Ph.D. degree in telecommunications from the KTH Royal Institute of Technology, Sweden, in 2011.

From 2012 to 2014, he held a joint post-doctoral position at the Alcatel-Lucent Chair on Flexible Radio, SUPELEC, France, and the KTH Royal Institute of Technology. He joined Linköping University, Sweden, in 2014, where he is currently an Associate Professor and a Docent with the Division of Communication Systems. He has authored the textbooks *Optimal Resource Allocation in Coordinated Multi-Cell Systems* in 2013 and *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency* in 2017. He is dedicated to reproducible research and has made a large amount of simulation code publicly available. He performs research on MIMO communications, radio resource allocation, machine learning for communications, and energy efficiency.

Dr. Björnson was a recipient of the 2014 Outstanding Young Researcher Award from the IEEE ComSoc EMEA, the 2015 Ingvar Carlsson Award, the 2016 Best Ph.D. Award from EURASIP, the 2018 IEEE Marconi Prize Paper Award in Wireless Communications, the 2019 EURASIP Early Career Award, and the 2019 IEEE Communications Society Fred W. Ellersick Prize. He also coauthored articles that received the Best Paper Awards at the conferences, including WCSP 2009, the IEEE CAMSAP 2011, the IEEE WCNC 2014, the IEEE ICC 2015, WCSP 2017, and the IEEE SAM 2014. Since 2017, he has been on the Editorial Board of the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING since 2016. He has performed MIMO research for more than ten years and has filed more than ten MIMO related patent applications.

Luca Sanguinetti (SM'15) received the Laurea Telecommunications Engineer degree *(cum laude)* and the Ph.D. degree in information engineering from the University of Pisa, Italy, in 2002 and 2005, respectively.

In 2004, he was a Visiting Ph.D. Student at the German Aerospace Center (DLR), Oberpfaffenhofen, Germany. From June 2007 to June 2008, he was a Postdoctoral Associate with the Department of Electrical Engineering, Princeton University. From July 2013 to October 2017, he was with Large Systems and Networks Group (LANEAS), CentraleSupélec, France. He is currently an Associate Professor with the 'Dipartimento di Ingegneria dell'Informazione', University of Pisa. He has coauthored the textbook *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency* in 2017. His expertise and general interests span the areas of communications and signal processing.

Dr. Sanguinetti was a recipient of the 2018 Marconi Prize Paper Award in Wireless Communications and coauthored an article that received the Young Best Paper Award from the ComSoc/VTS Italy Section. He was the recipient of the FP7 Marie Curie IEF 2013 "Dense deployments for green cellular networks". He was also a co-recipient of the two best conference paper awards: IEEE WCNC 2013 and IEEE WCNC 2014. He served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE JOURNAL ON SELECTED AREAS OF COMMUNICATIONS (series on Green Communications and Networking) and as a Lead Guest Editor for the IEEE JOURNAL ON SELECTED AREAS OF COMMUNICATIONS Special Issue on "Game Theory for Networks". He is currently serving as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, the IEEE TRANSACTIONS ON COMMUNICATIONS. He is also a member of the Executive Editorial Committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.