Reduced-complexity downlink cell-free mmWave Massive MIMO systems with fronthaul constraints

Guillem Femenias and Felip Riera-Palou
Mobile Communications Group - Universitat de les Illes Balears - 07122 Mallorca (Illes Balears), Spain
Email: {guillem.femenias,felip.riera}@uib.cat

***Abstract*—Cell-free architectures have recently emerged as a promising architecture with the potential to offer equal user-rates throughout the coverage area. Given the spectral congestion at sub-6 GHz bands, there is a pressing interest in evaluating the cell-free performance in the mmWave regime. This paper addresses the design and performance evaluation of the downlink segment of cell-free mmWave Massive MIMO system using hybrid precoders under the realistic assumption of capacity-constrained fronthaul links. Towards this end, a hybrid digital-analog beamforming is proposed where the high-dimensional analog part only depends on second-order large scale information. The low-dimensional digital part can then be implemented using standard precoding techniques that rely on instantaneous CSI. Numerical results demonstrate that this reduced-complexity architecture, when combined with an adequate user selection (scheduling), attains excellent Max-Min performance when operating under limited-fronthaul constraints.**

I. Introduction

Very recently, the underpinning idea of massive multiple-input multiple-output (MIMO) has been combined with the deployment of ultradense networks (UDN) giving rise to the concept of *cell-free massive MIMO* networks [1]. In these networks, a massive number of access points (APs) connected to a central processing unit (CPU) are distributed across the coverage area to coherently serve a large number of mobile stations (MSs) over the same time/frequency resources. Interestingly, using simple linear signal processing schemes, *cell-free massive MIMO* claims to provide uniformly good quality of service (QoS) to the whole set of served MSs throughout the entire coverage area.

Since the microwave radio spectrum is highly congested, the so-called millimeter wave (mmWave) bands [2] have recently attracted the attention of the research community. The very small wavelengths of mmWaves, combined with the technological advances in low-power CMOS radio frequency (RF) miniaturization, allow for the integration of a large number of antenna elements into small form factors. Large antenna arrays can then be used to effectively implement mmWave massive MIMO schemes that, with appropriate beamforming, can more than compensate for the orders-of-magnitude increase in free-space path-loss produced by the use of higher frequencies. The performance of cell-free massive MIMO using conventional sub-6 GHz frequency bands under infinite- or finite-capacity fronthaul links has been extensively studied in, for instance, [1], [3]–[5] but always assuming the use of fully digital precoders, a solution difficult to replicate in practice in the mmWave range. In fact, most mmWave systems rely on hybrid digital-analog signal processing architectures typically implemented using analog phase shifters and/or analog switches for the RF front-end in conjunction with low-dimensional baseband digital precoders [6]. Despite its evident potential, as far as we know, besides [7], [8] there is no other research work on cell-free mmWave massive MIMO systems and, furthermore, the authors of these works did not consider the fact that these systems require of a substantial information exchange between the APs and the CPU via capacity-constrained fronthaul links. Moreover, they also considered the use of oversimplified mmWave channel models and RF precoding stages, without constraining the available number of RF-chains at each AP.

Our main aim in this paper is to address the design and performance evaluation of realistic cell-free mmWave massive MIMO systems using hybrid precoders and assuming the availability of capacity-constrained fronthaul links connecting the APs and the CPU. In particular, the performance of the downlink (DL) of a cell-free mmWave massive MIMO system is considered by posing and solving max-min fairness resource allocation problems that take into account the effects of imperfect channel estimation, power control, non-orthogonality of pilot sequences, and fronthaul capacity constraints. Along the way, a hybrid beamforming implementation is proposed where the high-dimensionality RF stage is based on large-scale second-order statistics of the channel while limiting the use of instantaneous CSI to the low-dimensional baseband multiuser-MIMO (MU-MIMO) precoding/decoding stages.

II. System model

Let us consider the downlink of a cell-free massive MIMO system where a CPU coordinates the communication between M APs and K single-antenna MSs randomly distributed in a large area. Each of the APs communicates with the CPU via error-free fronthaul links with DL capacity C_{Fd}. Baseband processing of the transmitted/received signals is performed at the CPU, while the RF operations are carried out at the APs. Each AP is equipped with an array of $N > K$ antennas and $L \leq N$ RF chains. A fully-connected architecture is considered where each RF chain is connected to the whole set of antenna elements using N analog phase shifters [6]. Without loss of generality, it is assumed in this paper that the number of active RF chains at each of the APs in the network is equal to $L_A = \min\{K, L\}$. That is, if $K \leq L$, all APs in the cell-free network provide service to the whole MS set whereas if $K > L$, each AP can only provide service to L out

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

of the K MSs in the network, thus requiring of an algorithm to decide which are the MSs to be served from each APs.

The propagation channels linking the APs to the MSs are typically characterized by small-scale parameters that are (almost) static over a coherence time-frequency interval of τ_c time-frequency samples (see [9, Chapter 2]), and large-scale parameters (i.e., path loss propagation losses and covariance matrices) that can be safely assumed to be static over a time-frequency interval $\tau_{Lc} \gg \tau_c$. As shown in the following subsections, these channel characteristics can be leveraged to simplify both the channel estimation and the precoding/combining processes. In particular, DL and uplink (UL) transmissions between APs and MSs are organized in a half-duplex time division duplexing (TDD) operation whereby each coherence interval is split into three phases, namely, the UL training phase, the DL payload data transmission phase and the UL payload data transmission phase, and every *large-scale coherence interval* τ_{Lc} the system performs an estimation of the large-scale parameters of the channel. As typically done in a Massive MIMO TDD-based context, channel reciprocity is assumed to hold, thus avoiding the need for downlink training [9]. Unlike sub-6 GHz channels, mmWave propagation is characterized by very high distance-based propagation losses that lead to sparse scattering multipath propagation. Furthermore, the use of mmWave transmitters and receivers with large tightly-packet antenna arrays results in high antenna correlation levels. In this work, use is made of the discrete-time narrowband clustered channel model proposed by Akdeniz *et al.* in [10] and further extended in [11]. Under this model, when a link is not in outage is characterized using a standard pathloss model with shadowing given by

$$\mathrm{PL}(d_{mk})[dB] = \alpha + 10\beta \log_{10}(d_{mk}) + \chi_{mk}, \tag{1}$$

where α and β are the least square fits of floating intercept and slope and depend on the carrier frequency and on whether the link is in line-of-sight (LOS) or non-line-of-sight (NLOS) (see [10, Table I]). Parameter χ_{mk} denotes the large-scale shadow fading component, which is modelled as a zero mean spatially correlated normal random variable with standard deviation σ_χ (again, see [10, Table I] to obtain the typical values of σ_χ for LOS and NLOS links) whose spatial correlation model is described in [1, (54)-(55)].

The channel vector $\boldsymbol{h}_{mk} \in \mathbb{C}^{N\times 1}$ between AP m and MS k will be modelled as the sum of the contributions of C_{mk} scattering clusters, each contributing P_{mk} propagation paths

$$\boldsymbol{h}_{mk} = \sum_{c=1}^{C_{mk}} \sum_{p=1}^{P_{mk}} \alpha_{mk,cp} \boldsymbol{a}\left(\theta_{mk,cp}, \phi_{mk,cp}\right), \tag{2}$$

where $\alpha_{mk,cp}$ is the complex small-scale fading gain on the pth path of cluster c, and $\boldsymbol{a}\left(\theta_{mk,cp}, \phi_{mk,cp}\right)$ represents the AP normalized array response vector at the azimuth and elevation angles $\theta_{mk,cp}$ and $\phi_{mk,cp}$, respectively (see [10, Section III.E]). For notational convenience, we define the $N \times K$ matrix $\boldsymbol{H}_m = [\boldsymbol{h}_{mk} \dots \boldsymbol{h}_{mK}]$ as the matrix containing the channel responses from AP m to all the scheduled K users.

Although the small-scale fading gains $\alpha_{mk,cp}$ are considered static throughout the coherence interval and then change independently (i.e., block fading), the spatial covariance matrices $\boldsymbol{R}_{mk} = \mathbb{E}\left\{\boldsymbol{h}_{mk}\boldsymbol{h}_{mk}^H\right\}$ are assumed to vary at a much slower pace (i.e., $\tau_{Lc} \gg \tau_c$). Using spatial channel covariance estimation for hybrid analog-digital MIMO architectures (e.g. [12]), it can be safely assumed that $\boldsymbol{R}_{mk}$ is known at the corresponding mth AP.

III. Transmitter processing

A. RF precoder design

Using eigen-decomposition, the covariance matrix of the propagation channel linking MS k and AP m can be expressed as $\boldsymbol{R}_{mk} = \boldsymbol{U}_{mk}\boldsymbol{\Lambda}_{mk}\boldsymbol{U}_{mk}^H$, where $\boldsymbol{\Lambda}_{mk} = \mathrm{diag}\left([\lambda_{mk,1} \dots \lambda_{mk,r_{mk}}]\right)$ contains the r_{mk} non-null eigenvalues of $\boldsymbol{R}_{mk}$, and $\boldsymbol{U}_{mk}$ is the $N \times r_{mk}$ matrix of the corresponding eigenvectors. Hence, assuming the use of (constrained) statistical eigen beamforming [13], the analog RF precoder/combiner can be designed as

$$\boldsymbol{W}_m^{RF} = \left[\boldsymbol{w}_{m\kappa_{m1}}^{RF} \quad \dots \quad \boldsymbol{w}_{m\kappa_{mL_A}}^{RF}\right] \tag{3}$$

where $\boldsymbol{w}_{m\kappa_{ml}}^{RF} = e^{-j\angle \boldsymbol{u}_{m\kappa_{ml},\max}}$, $\boldsymbol{u}_{mk,\max}$ is the dominant eigenvector of $\boldsymbol{R}_{mk}$ associated to the maximum eigenvalue $\lambda_{mk,\max}$, and the function $\angle\boldsymbol{x}$ returns the phase angles, in radians, for each element of the complex vector $\boldsymbol{x}$. Note that using the RF precoding/combining matrix, the equivalent channel vector between MS k and AP m, including the RF precoding/decoding matrix, is defined as

$$\boldsymbol{g}_{mk} = \boldsymbol{W}_m^{RF^T} \boldsymbol{h}_{mk} \ \in \ \mathbb{C}^{L_A \times 1}, \tag{4}$$

whose dimension is much less than the number of antennas of the massive MIMO array used at the mth AP, thus largely simplifying the small-scale training phase.

B. Channel estimation

Communication in any coherence interval of a TDD-based massive MIMO system invariably starts with the MSs sending the pilot sequences to allow the channel to be estimated at the APs. Let τ_p denote the UL training phase duration (measured in samples on a time-frequency grid) per coherence interval. During the UL training phase, all K MSs simultaneously transmit pilot sequences of τ_p samples to the APs resulting in the $L_A \times \tau_p$ received UL signal matrix at the mth AP

$$\boldsymbol{Y}_{p_m} = \sqrt{\tau_p P_p} \sum_{k'=1}^{K} \boldsymbol{g}_{mk'} \boldsymbol{\varphi}_{k'}^T + \boldsymbol{N}_{p_m}, \tag{5}$$

where P_p is the transmit power of each pilot symbol, $\boldsymbol{\varphi}_k$ denotes the $\tau_p \times 1$ training sequence assigned to MS k, with $\|\boldsymbol{\varphi}_k\|_F^2 = 1$, and $\boldsymbol{N}_{p_m}$ is an $L_A \times \tau_p$ matrix of i.i.d. additive noise samples with each entry distributed as $\mathcal{CN}(0, \sigma_u^2(N))$. Ideally, training sequences should be chosen to be mutually orthogonal, however, since in most practical scenarios it holds that $K > \tau_p$, a given training sequence is assigned to more than one MS, thus resulting in the so-called pilot contamination. Channel estimation is conducted adhering to the minimum mean square error (MMSE) criterion, resulting in a estimated channel vector [1]

$$\hat{\boldsymbol{g}}_{mk} = \sqrt{\tau_p P_p} \boldsymbol{R}_{mk}^{RF} \boldsymbol{Q}_{mk}^{-1} \boldsymbol{Y}_{p_m} \boldsymbol{\varphi}_k^*, \tag{6}$$

where

$$\boldsymbol{R}_{mk}^{RF} = \mathbb{E}\left\{\boldsymbol{g}_{mk}\boldsymbol{g}_{mk}^H\right\} = \boldsymbol{W}_m^{RF^T} \boldsymbol{R}_{mk} \boldsymbol{W}_m^{RF^*}, \tag{7}$$

and $\boldsymbol{Q}_{mk} = \tau_p P_p \sum_{k'=1}^{K} \boldsymbol{R}_{mk'}^{RF} \left|\boldsymbol{\varphi}_{k'}^T \boldsymbol{\varphi}_k^*\right|^2 + \sigma_u^2(N)\boldsymbol{I}_{L_A}$.

The MMSE channel vector estimates can be shown to be distributed as

$$\hat{\boldsymbol{g}}_{mk} \sim \mathcal{CN}\left(\boldsymbol{0}, \hat{\boldsymbol{R}}_{mk}^{RF}\right) \text{ with } \hat{\boldsymbol{R}}_{mk}^{RF} \triangleq \tau_p P_p \boldsymbol{R}_{mk}^{RF} \boldsymbol{Q}_{mk}^{-1} \boldsymbol{R}_{mk}^{RF^H}.$$

Furthermore, the channel vector $\boldsymbol{g}_{mk}$ can be decomposed as $\boldsymbol{g}_{mk} = \hat{\boldsymbol{g}}_{mk} + \tilde{\boldsymbol{g}}_{mk}$, where $\tilde{\boldsymbol{g}}_{mk}$ is the MMSE channel estimation error, which is statistically independent of both $\boldsymbol{g}_{mk}$ and $\hat{\boldsymbol{g}}_{mk}$. For notational convenience, matrices $\hat{\boldsymbol{G}_m} = [\hat{\boldsymbol{g}}_{m1}, \ldots, \hat{\boldsymbol{g}}_{mK}]$ and $\tilde{\boldsymbol{G}_m} = [\tilde{\boldsymbol{g}}_{m1}, \ldots, \tilde{\boldsymbol{g}}_{mK}]$ correspond to the estimated and error matrices for AP m.

C. Baseband precoder design

Relying on the MMSE estimates of the equivalent baseband channel $\hat{\boldsymbol{g}}_{mk}$, and in line with [3], we can define the classical zero-forcing (ZF) MU-MIMO baseband precoder as $\boldsymbol{W}^{BB} = \hat{\boldsymbol{G}}^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1}$ or, equivalently,

$$\boldsymbol{W}_m^{BB} = \hat{\boldsymbol{G}}_m^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \quad \forall m, \tag{8}$$

where $\boldsymbol{G} = [\boldsymbol{G}_1^T \ldots \boldsymbol{G}_M^T]^T$, with $\boldsymbol{G}_m = \boldsymbol{W}_m^{RF^T} \boldsymbol{H}_m$, representing the equivalent (RF precoded) MIMO channel matrix between the K MSs and the M APs.

D. Quantized downlink data transmission

Let us define $\boldsymbol{s} = [s_1 \ldots s_K]^T$ as the $K \times 1$ vector of symbols to be conveyed to the MSs, holding $E\left\{\boldsymbol{s}\boldsymbol{s}^H\right\} = \boldsymbol{I}_K$. Let us also define $\boldsymbol{x}_m = \mathcal{P}_m(\boldsymbol{s})$ as the $N \times 1$ vector of signals transmitted from the mth AP, where $\mathcal{P}_m(\boldsymbol{s})$ is used to denote the mathematical operations (linear and/or non-linear) used to obtain $\boldsymbol{x}_m$ from $\boldsymbol{s}$. Note that this vector must comply with a power constraint $\mathbb{E}\left\{\|\boldsymbol{x}_m\|_F^2\right\} \leq \overline{P}_m$, where $\overline{P}_m$ is the maximum average transmit power available at AP m.

In particular, the processing of symbol vector $\boldsymbol{s}$, includes, first, a baseband precoding task at the CPU, second, a compressing process of the data that must be sent from the CPU to the APs through the fronthaul links and, third, an RF precoding task at each of the APs. Using [14], the distortion introduced by the quantization/unquantization processes for the signals processed at the mth AP, can be modelled as $\hat{\mathcal{Q}}_m(\boldsymbol{x}) \triangleq \mathcal{Q}_m^{-1}(\mathcal{Q}_m(\boldsymbol{x})) = \boldsymbol{x} + \boldsymbol{q}_m$ with $\boldsymbol{q}_m$ denoting the quantization noise. Transmitter processing can be defined as

$$\begin{aligned} \boldsymbol{x}_m = \mathcal{P}_m(\boldsymbol{s}_d) &= \boldsymbol{W}_m^{RF} \hat{\mathcal{Q}}_m\left(\boldsymbol{W}_m^{BB} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s}\right) \\ &= \boldsymbol{W}_m^{RF}\left(\boldsymbol{W}_m^{BB} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s} + \boldsymbol{q}_m\right), \end{aligned} \tag{9}$$

where $\boldsymbol{W}^{BB} = \left[\boldsymbol{W}_1^{BB^T} \ldots \boldsymbol{W}_M^{BB^T}\right]^T \in \mathbb{C}^{ML_A \times K}$, with $\boldsymbol{W}_m^{BB} \in \mathbb{C}^{L_A \times K}$ denoting the baseband precoding matrix affecting the signal transmitted by the mth AP, and $\boldsymbol{\Upsilon} = \text{diag}\left([\upsilon_1 \ldots \upsilon_K]\right)$ is a $K \times K$ diagonal matrix containing the power control coefficients in its main diagonal, which should be chosen adhering to the mth AP's power constraint. According to results in [14], [15], quantization noise can be assumed to be statistically distributed as $\boldsymbol{q}_m \sim \mathcal{CN}\left(\boldsymbol{0}, \sigma_{q_m}^2 \boldsymbol{I}\right)$. As shown in [14], this assumption is supported by the fact that large-block lattice quantization codes are able to approximate a Gaussian quantization noise distribution.

Based on results in [15], it can be shown that the required average rate at the mth AP, $\hat{C}_m$, to transfer the quantized vector $\hat{\mathcal{Q}}_m\left(\boldsymbol{W}_m^{BB} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s}\right)$ on the corresponding DL fronthaul link (in bps/Hz) can be upper-bounded by

$$\hat{C}_m \leq \log_2 \det\left(\frac{1}{\sigma_{q\,m}^2} \sum_{k=1}^{K} \upsilon_k \boldsymbol{R}_{mk}^{BB} + \boldsymbol{I}_{L_A}\right), \tag{10}$$

where $\boldsymbol{R}_{mk}^{BB} = \mathbb{E}\left\{\boldsymbol{w}_{mk}^{BB} \boldsymbol{w}_{mk}^{BB^H}\right\}$.

Using the proposed hybrid compress-after-precoding (CAP) approach, and denoting by $\boldsymbol{g}_k, \hat{\boldsymbol{g}}_k$ and $\tilde{\boldsymbol{g}}_k$ the kth row of $\boldsymbol{G}, \hat{\boldsymbol{G}}$ and $\tilde{\boldsymbol{G}}$, respectively, the signal received by the kth MS follows

$$\begin{aligned} y_k =& \boldsymbol{g}_k^T \hat{\boldsymbol{G}}^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s} + \eta_k \\ =& \left(\hat{\boldsymbol{g}}_k^T + \tilde{\boldsymbol{g}}_k^T\right) \hat{\boldsymbol{G}}^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s} + \eta_k \\ =& \sqrt{\upsilon_k} s_k + \tilde{\boldsymbol{g}}_k^T \hat{\boldsymbol{G}}^* \left(\hat{\boldsymbol{G}}^T \hat{\boldsymbol{G}}^*\right)^{-1} \boldsymbol{\Upsilon}^{1/2} \boldsymbol{s} + \eta_k \end{aligned} \tag{11}$$

where $\eta_k = \boldsymbol{g}_k^T [\boldsymbol{q}_1^T, \ldots, \boldsymbol{q}_M^T]^T + n_k$ with $n_k \sim \mathcal{N}(0, \sigma^2)$. The first term denotes the useful received signal, the second term contains the interference due to the use of imperfect channel state information (CSI) (pilot contamination), and the third term encompass both the quantification and thermal noise.

IV. Max-min power allocation and quantization

Analysis techniques similar to those applied, for instance, in [1], [3], [9], can be used to derive the DL achievable rate. In particular, if the sum of the second and third terms on the right hand side (RHS) of (11), are treated as *effective noise*, the achievable rate (in bits/s/Hz) of MS k using the analog precoders $\boldsymbol{W}_m^{RF}$, for all $m \in \{1, \ldots, M\}$, and a ZF baseband precoder is given by $R_{dk} = \log_2\left(1 + \text{SINR}_{dk}\right)$, with

$$\text{SINR}_k = \frac{\upsilon_k}{\sum_{k'=1}^{K} \upsilon_{k'} \varpi_{kk'} + \sigma_{\eta_k}^2}, \tag{12}$$

where $\varpi_{kk'} = \left[\text{diag}\left(\mathbb{E}\left\{\boldsymbol{W}^{BB^H} \tilde{\boldsymbol{g}}_k^* \tilde{\boldsymbol{g}}_k^T \boldsymbol{W}^{BB}\right\}\right)\right]_{k'}$ and $\sigma_{\eta_k}^2 = \sum_{m=1}^{M} \sigma_{q\,m}^2 \,\text{tr}\left(\boldsymbol{R}_{mk}^{RF}\right) + \sigma^2$.

In line with previous research works [1], [3], power control coefficients υ_k, for all $k \in \{1, \ldots, K\}$, and the quantization noise variances $\sigma_{q\,dm}^2$, are sought for all $m \in \{1, \ldots, M\}$, that maximize the minimum of the achievable DL rates of all MSs while satisfying the average transmit power and DL

TABLE I: Summary of default simulation parameters

Parameters	Value
Carrier frequency: f_0	28 GHz
Bandwidth: B	20 MHz
Side of the square coverage area: D	200 m
AP/MS antenna heights: h_{AP}/h_{MS}	15 m/1.65 m
Noise figure at the MS: NF_{MS}	9 dB
Available average power at the AP: $\overline{P}_m$	200 mW
Coherence interval length: τ_c	200 samples
Training phase length: τ_p	15 samples

fronthaul capacity constraints at each AP. Mathematically, this optimization problem can be formulated as

$$\begin{aligned}
&\max_{\substack{\boldsymbol{\Upsilon}\succeq 0\\ \boldsymbol{\sigma}_q\succeq 0}}\ \min_{k\in\{1,\dots,K\}} \frac{\upsilon_k}{\sum_{k'=1}^{K}\upsilon_{k'}\varpi_{kk'}+\sigma^2_{\eta_k}}\\
&\text{s.t.}\ \sum_{k=1}^{K}\upsilon_k\theta_{mk}^{BB/RF}\leq \overline{P}_m-\sigma^2_{q_m}L_A N,\ \forall m, \qquad (13)\\
&\quad \log_2\det\left(\sum_{k=1}^{K}\frac{\upsilon_k}{\sigma^2_{q_m}}\boldsymbol{R}_{mk}^{BB}+\boldsymbol{I}_{L_A}\right)\leq C_{Fd},\ \forall m,
\end{aligned}$$

using the definitions $\boldsymbol{\sigma}_q = [\sigma_{q_1}\ldots\sigma_{q_M}]^T$ and $\theta_{mk}^{BB/RF} = \mathbb{E}\left\{\left\|\boldsymbol{W}_m^{RF}\boldsymbol{w}_{mk}^{BB}\right\|_F^2\right\}$, and the fact that $\left\|\boldsymbol{W}_m^{RF}\right\|_F^2 = L_A N$. Optimization problem (13) is characterized by a continuous objective and constraint functions of interdependent block variables, namely, $\boldsymbol{\Upsilon}$ and $\boldsymbol{\sigma}_q$, which can be solved using the so-called block coordinate descend (BCD) method. In BCD, at each iteration and in a cyclic order, one of the blocks is optimized while the remaining variables are held fixed [16]. Convergence of the BCD method is ensured whenever each of the subproblems to be optimized in each iteration can be exactly solved to its unique optimal solution. As shown in [17], problem (13) can be transformed into an equivalent convergent quasi-linear optimization problem that can be solved using conventional standard convex optimization methods [1], [3].

V. Numerical results

Simulations results are now presented to quantitatively study the performance of the proposed cell-free mmWave massive MIMO network with constrained-capacity fronthaul links. For simplicity of exposition, and without loss of generality, a cell-free scenario is considered where the M APs and K MSs are uniformly distributed at random within a square coverage area of size $D \times D$ m^2. A modified version of the discrete-time narrowband clustered channel model proposed by Akdeniz *et al.* in [10, Table I] is used in the performance evaluation. Furthermore, as in [1], a shadow fading spatial correlation model with two components is also considered where the decorrelation distance is set to $d_{decorr} = 50$ m and the parameter δ is set to 0.5. For scenarios where $K > \tau_p$, pilot allocation is conducted using dissimilarity cluster-based pilot assignment (DCPA) trying to minimize the effects of pilot contamination (see [17] for details). User scheduling, enforced whenever $K > L$, is conducted using an iterative-reverse algorithm (similar to that used in graph theory to construct a minimum spanning tree from a edge-weighted graph) proposed in [17]. The rest of simulation parameters are summarized in Table I. The max-min achievable rate per user is plotted on the left side of Fig. 1 against the number of active MSs in the network, assuming the use of different fronthaul capacities. Note that for the network setups under consideration, using fronthaul links with a capacity of 256 bit/s/Hz is virtually equivalent to using infinite-capacity fronthauls. As expected, results show that increasing the fronthaul capacity is always beneficial if the main aim is to increase the achievable max-min user rate. Nevertheless, it is worth stressing that, keeping all the other parameters constant, the marginal increment of performance produced by each new increment of the fronthaul capacity suffers from the law of diminishing returns, especially for network setups with a high number of active MSs. In particular, is hardly justifiable increasing the fronthaul capacity beyond 64 bit/s/Hz. The right plot of Fig. 1 depicts the achievable max-min user rate against the number of active MSs when considering different number of antenna elements (note that the number of RF chains remains fixed to $L = 8$). Notice how irrespective of the number of active MSs in the cell-free network, increasing the number of antenna elements at the APs in scenarios with high capacity fronthaul links ($C_{Fd} = 64$ bit/s/Hz), although moderate and again subject to the law of diminishing returns, always produces an increase in the achievable max-min user rate. Note in both plots the considerable performance drop caused by pilot contamination when the number of users in the system exceeds the pilot sequence length (i.e., $K > 15$).

In order to deepen in the study of the impact the RF infrastructure may have, the average max-min user rate is plotted in Fig. 2 against the number of antenna elements (left plot) and RF chains (right plot), respectively, for different values of the fronthaul capacities and assuming a fixed number of $K = 20$ active MSs. In network setups using very high capacity fronthaul links (i.e., $C_{Fd} = 256$ bit/s/Hz), increasing the number of antenna elements N and/or the number of RF chains L (up to $L = K$) is always beneficial as, in this case, the noise introduced by the quantization process is negligible and the system can take full advantage of the increased RF resources. As the capacity of the fronthaul links decreases, however, the amount of noise introduced by the quantization process increases with both N and L and, therefore, a situation arises where the potential performance improvement provided by the increase of N and/or L is compromised by the performance reduction due to fronthaul capacity constraints. On the one hand, it can be observed on the left plot in Fig. 2 that there is a certain fronthaul capacity constraint value (near 24 bit/s/Hz in this case) under which increasing the number of antenna elements at the array is counterproductive. On the other hand, results presented on the right plot in Fig. 2 show that, for fixed K and N, there is always an optimal number of RF chains to be deployed (or activated) at the APs that dependens on the fronthaul capacity. In this scenario, the optimal number of RF chains become $L = 10$, 4, and 1 when

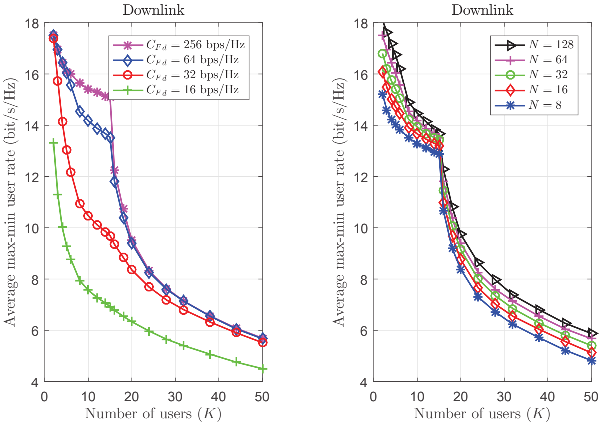

Fig. 1: Average max-min rate per user *versus* the number of active MSs for different values of the fronthaul capacities (left, with $N = 64$ antennas) and different number of AP antennas (right, with $C_f = 64$ bits/s/Hz). For both plots, $L = 8$ RF chains.

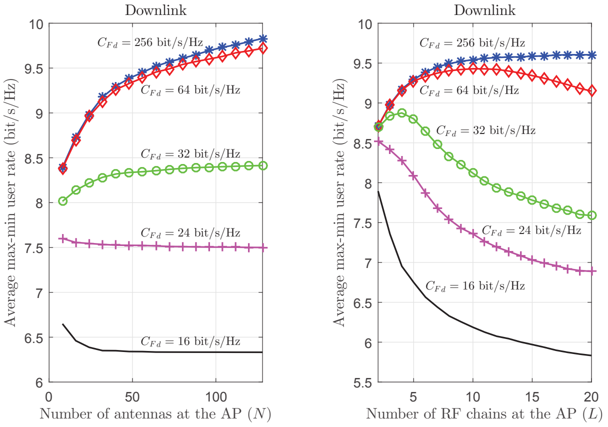

Fig. 2: Average max-min rate per user *versus* number of antennas at the APs (left, with $L = 8$ RF chains) and *versus* number of RF chains at the APs (right, with $N = 64$ antennas) for different values of the fronthaul capacities. For both plots $K = 20$ users.

using fronthaul capacities of 64 bit/s/Hz, 32 bit/s/Hz and less than 24 bit/s/Hz, respectively.

VI. Conclusion

A novel analytical framework targeting cell-free mmWave massive MIMO networks using capacity-constrained fronthaul links has been presented. The proposed framework considers the use of low-complexity hybrid precoders/decoders and the use of large-block lattice quantization codes able to approximate a Gaussian quantization noise distribution. Max-min power allocation and fronthaul quantization optimization problems have been posed thanks to the development of mathematically tractable expressions for both the per-user achievable rates and the fronthaul capacity consumption. Results show that, although increasing the fronthaul capacity and/or the density of APs per area unit is always beneficial from the point of view of the achievable max-min user rate, the marginal increment of performance produced by each new increment of these parameters suffers from the law of diminishing returns, especially for network setups with a high number of active MSs. Moreover, simulation results indicate that, as the capacity of the fronthaul link decreases, the potential performance improvement provided by the increase of the number of antenna elements N and/or the number of RF chains L is compromised by the performance reduction due to the corresponding increase of the fronthaul quantization noise.

Acknowledgements: Work supported by MINECO under project TERESA-TEC2017-90093-C3-3-R (AEI/FEDER,UE), Spain.

References

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

[2] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks - with a focus on propagation models," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6213–6230, Dec 2017.

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

[4] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, and M. Debbah, "Cell-free massive MIMO with limited backhaul," *arXiv preprint arXiv:1801.10190*, 2018.

[5] M. N. Boroujerdi, A. Abbasfar, and M. Ghanbari, "Cell free massive MIMO with limited capacity fronthaul," *Wireless Personal Communications*, October 2018.

[6] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 134–141, Sept 2017.

[7] M. Alonzo and S. Buzzi, "Cell-free and user-centric massive MIMO at millimeter wave frequencies," in *IEEE PIMRC*, Oct 2017, pp. 1–5.

[8] M. Alonzo, S. Buzzi, and A. Zappone, "Energy-efficient downlink power control in mmwave cell-free and user-centric massive MIMO," *CoRR*, vol. abs/1805.05177, 2018. [Online]. Available: http://arxiv.org/abs/1805.05177

[9] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of massive MIMO*. Cambridge University Press, 2016.

[10] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[11] M. K. Samimi and T. S. Rappaport, "Ultra-wideband statistical channel model for non line of sight millimeter-wave urban channels," in *2014 IEEE Global Communications Conference*, Dec 2014, pp. 3483–3489.

[12] A. Adhikary, E. A. Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch, "Joint spatial division and multiplexing for mm-Wave channels," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1239–1255, June 2014.

[13] R. Mai, T. Le-Ngoc, and D. H. N. Nguyen, "Two-timescale hybrid RF-baseband precoding with MMSE-VP for multi-user massive MIMO broadcast channels," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4462–4476, July 2018.

[14] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Transactions on Information Theory*, vol. 42, no. 4, pp. 1152–1159, Jul 1996.

[15] G. Femenias and F. Riera-Palou, "Multi-layer downlink precoding for cloud-RAN systems using full-dimensional massive MIMO," *IEEE Access*, vol. 6, pp. 61 583–61 599, 2018.

[16] A. Beck and L. Tetruashvili, "On the convergence of block coordinate descent type methods," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2037–2060, 2013.

[17] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity," *IEEE Access*, vol. 7, pp. 44 596–44 612, 2019.