IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 16, 2020, accepted June 19, 2020, date of publication June 23, 2020, date of current version July 2, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3004499*

# Fronthaul-Constrained Cell-Free Massive MIMO With Low Resolution ADCs

**GUILLEM FEMENIAS, (Senior Member, IEEE), AND FELIP RIERA-PALOU, (Senior Member, IEEE)**
Mobile Communications Group, University of the Balearic Islands, 07122 Palma, Spain

Corresponding author: Guillem Femenias (guillem.femenias@uib.es)

This work was supported in part by the Agencia Estatal de Investigación and Fondo Europeo de Desarrollo Regional (AEI/FEDER, UE), Ministerio de Economía y Competitividad (MINECO), Spain, through the TERESA Project, under Grant TEC2017-90093-C3-3-R.

**ABSTRACT** In cell-free massive MIMO networks, a large number of distributed access points (APs) provide service to a much smaller number of mobile stations (MSs) over the same time/frequency resources. The key idea is to use a central processing unit (CPU) to manage such a densely populated network of APs. This centralization helps reducing operational costs and eases implementation of joint power control and coherent signal processing through a proper orchestration of the functional split between the CPU and the APs. Cell-free massive MIMO networks, however, are often subject to stringent capacity requirements on the fronthaul links connecting the APs to the CPU and thus, low-resolution ADCs must be used to quantize the signals shared among CPU and APs. In this paper, analytical closed-form expressions for the achievable user rates on both the uplink (UL) and downlink (DL) of a fronthaul-capacity constrained cell-free massive MIMO network using low-resolution ADCs are obtained. These expressions, jointly with the use of theoretical models characterizing the fronthaul capacity consumption of different CPU-AP functional splits, allow posing max-min fairness power control optimization problems that can be solved using standard convex optimization algorithms. Numerical results show that, under fronthaul capacity constraints, CPU-AP functional splits where the precoding/decoding schemes are implemented at the APs are clearly outperformed by those functional splits in which, thanks to sharing CSI among APs and CPU, the precoding/decoding functions are implemented at the CPU. In contrast, if the limiting factor is the resolution of the ADCs used to quantize the samples to be transmitted on the fronthaul links, the preferred CPU-AP functional splits are those in which the baseband processing is performed at the APs. Moreover, they also reveal that in such functional splits there is always an optimal range of values of the UL fronthaul capacity fraction allocated to share the CSI.

**INDEX TERMS** Cell-free massive MIMO, capacity-constrained fronthaul, normalized conjugate beamforming, matched filtering, CPU-AP functional split, low-resolution ADCs.

## I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) has recently emerged as one of the fundamental physical layer pillars of the so-called 5G and beyond-5G wireless networks [1], [2]. The underlying advantage of massive MIMO, compared to *classical* multi-user MIMO, is that it can provide very high spectral and energy efficiencies by relying on rather simple signal processing, without the need for any base station (BS) cooperation [3]. Although massive MIMO arrays at the BSs have been traditionally arranged in compact collocated setups, they can also be organized in spatially distributed configurations [1], [2], [4]. Distributed massive MIMO architectures are reminiscent of concepts such as distributed antenna system (DAS) [5], network MIMO [6], [7], coordinated multipoint (CoMP) transmission [8] or cloud radio access network (C-RAN) [9], but all these arrangements can be essentially considered as different incarnations of a cooperative cellular infrastructure. To the best of authors' knowledge, the first distributed massive MIMO architecture was proposed in [10] where the uplink segment was investigated when relying on maximum-ratio combining (MRC) in combination with a BS selection procedure that effectively limited the number of remote radio heads (RRHs) involved

The associate editor coordinating the review of this manuscript and approving it for publication was Martin Reisslein.

in the detection of the signal transmitted by a particular mobile station (MS). A similar setup was addressed in [11] but using minimum mean square error (MMSE) detection, which provides an upper bound on the performance linear detectors can offer in practical setups (i.e., finite number of RRHs). Conceptually similar to the C-RAN and distributed massive MIMO architectures, but explicitly renouncing to the cellular network philosophy, an alternative distributed massive MIMO-based infrastructure has been recently termed as *cell-free massive MIMO* [4], [12]. The underlying idea is that a massive number of access points (APs) distributed across the coverage area are connected to a central processing unit (CPU) and, as in the cellular collocated massive MIMO schemes, use very simple signal processing schemes to exploit the channel hardening and favorable propagation conditions to coherently serve a large number of MSs on the same time-frequency resource.

The distribution of antennas over a large area allows for an efficient exploitation of large-scale diversity while bringing network infrastructure physically closer to MSs to offer a much higher coverage probability than collocated massive MIMO architectures [1], [2], [4], [12], [13]. However, this comes at the cost of increased fronthaul capacity requirements. Despite capacity-constrained fronthaul links may dramatically influence the performance of cell-free massive MIMO networks, most research papers on this topic rely on the assumption of infinite-capacity fronthaul links (see, for instance, [4], [13]–[17]). Limited fronthaul effects have been previously considered in the context of CoMP using both, rate-distortion information theoretical arguments [18]–[20] or simple quantization mechanisms [21]. Interestingly, fronthaul limitations invariably bring along a new degree of freedom in the form of the functional split describing where and in what order are the precoding and quantizing operations implemented. Authors in [19], [20], [22] have recently shown that, in the downlink of a C-RAN network, the fronthaul capacity plays a decisive role in deciding the best functional split: whereas under mild fronthaul constraints it is advantageous to precode at the APs, CPU-based precoding is to be preferred when the fronthaul is severely constrained. Cell-free massive MIMO networks using capacity-constrained fronthaul links have only been recently considered, under very specific scenario-dependent conditions, in [23]–[27]. In particular, Bashar *et al.* in [23] only consider the uplink (UL) and, in addition, they assume the use of uniform quantizers with a fixed number of bits/sample. In [24], the authors solely focus on the downlink (DL) and assume the use of a specific CPU-AP functional split in which the control layer, in charge of delivering channel estimates to the CPU, is not subject to any capacity constraint, whereas the data layer, in charge of delivering user's precoded signals to the APs, is subject to capacity constrained links. Zhang *et al.* in [25] only evaluate the UL and, although they assume the use of low-resolution analog-to-digital converters (ADCs), do not consider the effects that the use of capacity constrained fronthaul links may have on the spectral efficiency of the network. Masoumi *et al.* in [26], analyze the performance provided by three different transmission strategies at the APs but, unfortunately, they do not assume the use of conventional low-resolution ADCs but the use of sophisticated compressors that, only by assuming the compression of very large signal sequences, can be modeled based on the rate-distortion theory. Finally, the cell-free massive MIMO scenario considered in [27] is specifically designed to exploit the characteristics of the so-called millimeter wave (mmWave) bands and massive MIMO architectures based on hybrid analog/digital zero-forcing (ZF) precoding/decoding schemes. Unfortunately, ZF strategies can only be implemented using specific CPU-AP functional splits, thus making this specific precoder unsuitable to analyse the effects that capacity-constrained fronthaul links may have on the different configurations. Our main aim in this paper is to fill in the gap left by previous research work on this topic by presenting a realistic general framework allowing a fair comparison between different CPU-AP functional splits in both the UL and DL of a fronthaul-constrained cell-free massive MIMO network using low-resolution ADCs. Specific contributions of this paper can be summarized as:

- As claimed by Chen and Björnson in [28], the levels of channel hardening and favourable propagation conditions provided by a cell-free massive MIMO network using conjugate beamforming (CB) strongly depend on the propagation environment, pathloss model and users' distribution. This hardening uncertainty casts doubts on the accuracy of the achievable rate expressions in such a cell-free massive MIMO setup as they can significantly understimate the true spectral efficiency. Fortunately, authors in [29] proposed a simple modification of the CB precoder, the normalized conjugate beamforming (NCB), targetting the DL of cell-free massive MIMO networks that can be easily shown to largely improve these hardening metrics. In contrast, for theUL segment, receiver matched filtering has been shown to outperform its normalized counterpart. Detailed mathematical analysis leads to approximate closed form expressions for the achievable rate, for both a matched filtering (MF)-based UL and an NCB-based DL, that are able to encompass the effects of the capacity-constrained fronthaul links. The results are also extended to consider diferent pilot allocation schemes.
- The quantizer noise model described by Mezghani and Nossek in [30], which is based on the Bussgang decomposition [31] and is sometimes referred to as the additive quantization noise model (AQNM) [32], is used to characterize the fronthaul bandwidth consumption of different CPU-AP functional splits, allowing us to provide a thorough comparison among them and discuss the impact they may have on the global performance of the network. Remarkably, as it will be shown over the next sections, CPU-AP functional splits performing the baseband signal processing operations at the

CPU exhibit far greater robustness against fronthaul bandwidth limitations in comparison to those where the processing is conducted at the APs.
- Using the mathematical models for both the achievable rates and the fronthaul bandwidth consumption, max-min per-user rate fairness power allocation and fronthaul quantization design strategies are devised that provide globally optimal solutions that can be solved using standard convex optimization algorithms. Unlike the seminal works on cell-free massive MIMO networking, the optimization problems posed and solved in this paper take into account the unavoidable fronthaul-capacity constraint. In particular, the resulting max-min optimization problems involve not only the power allocation coefficients but also the quantization impairments that a given capacity-constrained fronthaul link can support. Interestingly, both sets of optimization variables are shown to be deeply intertwined.

The remainder of this paper is organized as follows. In Section II we describe the proposed fronthaul-capacity constrained cell-free massive MIMO network. Different subsections are devoted to the description of the channel model, the quantizer model, the training phases, and the UL and DL payload transmission phases. Mathematical closed-form expressions for the achievable UL and DL user rates are derived in Section III and further developed in Appendices A and B. Section IV is dedicated to the characterization of the bandwidth consumption of both the UL and DL fronthaul links under different CPU-AP functional splits. Power allocation and quantization optimization processes are posed and solved in Section V. Numerical results and discussions are provided in Section VI and, finally, concluding remarks are summarized in Section VII.

*Notation:* Vectors and matrices are denoted by lower-case and upper-case boldface symbols. The $q$-dimensional identity matrix is represented by $\boldsymbol{I}_q$. The operators $\boldsymbol{X}^{-1}$, $\boldsymbol{X}^T$, $\boldsymbol{X}^*$ and $\boldsymbol{X}^H$ denote, respectively, the inverse, transpose, conjugate and conjugate transpose (also known as Hermitian) of matrix $\boldsymbol{X}$. The expectation operator is denoted by $\mathbb{E}\{\cdot\}$. Finally, $\mathcal{CN}(\boldsymbol{m}, \boldsymbol{R})$ denotes a complex Gaussian vector distribution with mean $\boldsymbol{m}$ and covariance $\boldsymbol{R}$, whose zero-mean part constructed by subtracting its mean is circularly symmetric, and $\mathcal{N}(0, \sigma^2)$ denotes a real valued zero-mean Gaussian random variable with standard deviation $\sigma$.

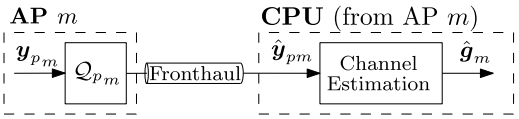
## II. SYSTEM MODEL

Following the original cell-free massive MIMO proposal in [4], this paper considers a wireless communications system where $M$ single antenna APs have been deployed on a large coverage area to simultaneously serve $K$ single antenna MSs on the same time-frequency resource. It is assumed that both APs and MSs are uniformly distributed over the coverage area and that all APs are connected to a CPU via fronthaul links with DL and UL capacities denoted by $C_{Fd}$ and $C_{Fu}$, respectively. As it is typically done in massive MIMO, DL and UL transmissions are organized in a time division duplex (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. In the UL training phase, all MSs transmit UL training pilots allowing the estimation of the propagation channels to every MS in the network.[1] Subsequently, these channel estimates are used to detect the signals transmitted from the MSs in the UL payload data transmission phase and to compute the precoding filters governing the DL payload data transmission.

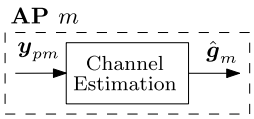
Although hybrid CPU-AP functional splits may be devised, only the two *classical* approaches will be considered in this paper, namely, the baseband processing at the CPU (BCU) and the baseband processing at the AP (BAP). The signal processing steps performed by these CPU-AP functional splits are schematically represented in the block diagrams shown in Fig. 1. In words (the mathematical details will be developed in the following subsections), these signal processing operations can be described as:

- **BCU-based CPU-AP functional split**: In the BCU-based approach, the received signal samples during the UL training phase (see Fig. 1(a)) are quantized at each AP and sent to the CPU, via the corresponding UL fronthaul links, where they are used for channel estimation and precoder/decoder design. During the UL payload data transmission phase (see Fig. 1(c)), the received samples at each AP are first quantized and then sent to the CPU, via the UL fronthaul link, where they are filtered for combining and detection. During the DL payload data transmission phase (see Fig. 1(e)), a precoded data signal is generated at the CPU for each of the MSs in the network and the $K$ precoded signals are then combined, quantized and sent to the APs, via the corresponding DL fronthaul links, where they are forwarded to the radio frequency (RF) chains for transmission. Note that in all these transmission phases, the channel state information (CSI) has been kept at the CPU.
- **BAP-based CPU-AP functional split**: In the BAP-based approach, in contrast, the received signal samples during the UL training phase (see Fig. 1(b)) are used at the APs to obtain a channel estimation and to calculate the baseband precoders and decoders. That is, the CSI is kept at the APs. During the UL payload data transmission phase (see Fig. 1(d)), the received samples at each AP are first filtered (decoded) to obtain $K$ signal samples corresponding to the $K$ active MSs in the network. These samples are then quantized, multiplexed and sent to the CPU, via the UL fronthaul link, where they are demultiplexed for combining and detection. During the DL payload data transmission phase (see Fig. 1(f)), the

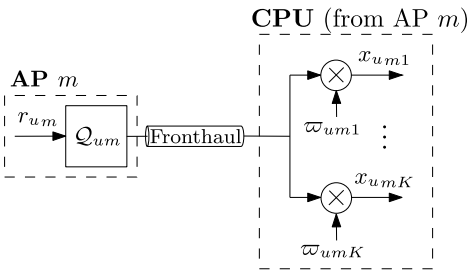
[1]Note that channel reciprocity can be exploited in TDD systems and therefore only UL pilots need to be transmitted. Furthermore, it is assumed that MSs do not need to estimate the effective channel gain as, due to massive MIMO channel hardening, this is very close to its expected value, a fairly easy to estimate deterministic constant.



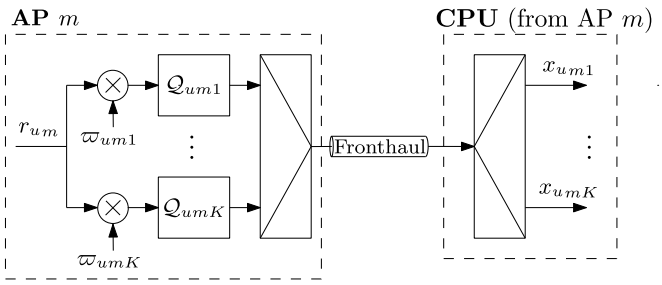

(a) BCU: Training phase



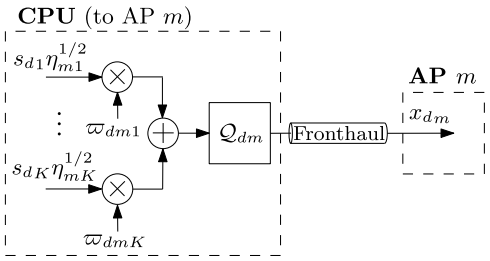

(b) BAP: Training phase



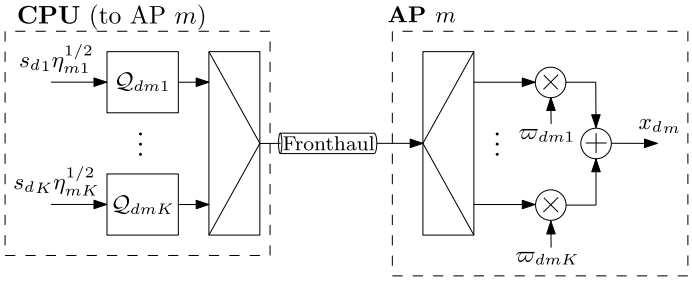

(c) BCU: UL payload data transmission phase




(d) BAP: UL payload data transmission phase




(e) BCU: DL payload data transmission phase




(f) BAP: DL payload data transmission phase

**FIGURE 1.** Schematic block diagram representing the operations performed by the BCU and BAP CPU-AP functional splits during the different TDD transmission phases.

power-controlled data symbols are quantized and multiplexed at the CPU and sent to the APs, via the corresponding DL fronthaul links. At the APs, they are precoded, combined and sent to the RF chains for transmission.

Note that the coefficients $\varpi_{umk}$ and $\varpi_{dmk}$, for all $m \in \{1,\ldots,M\}$ and $k \in \{1,\ldots,K\}$ are used to denote the MIMO decoding weights in the UL and MIMO precoding weights in the DL, respectively. These weighting coefficients are applied either at the CPU for the BCU-based functional split or at the APs for the BAP-based functional split. It is worth pointing out at this point that, in order to allow for a fair comparison between both functional splits, only rather simple MIMO signal processing techniques are considered in this work (see, for instance, [4], [29]). In particular, we focus on conjugate beamforming-based techniques suitable for distributed implementation that, under ideal/infinite fronthaul assumptions would render BCU and BAP equivalent. Note, however, that a potential advantage of the BCU-based functional split is that it enables the implementation of fully centralised processing in cell-free massive MIMO, thus opening the door to the implementation of MMSE data processing that, as pointed out in [33], may significantly outperform any form of conjugate beamforming, including the normalised conjugate beamforming proposed by Interdonato *et al.* in [29].

Critically, the combined duration/bandwidth of the training, and DL and UL payload transmission phases, denoted as $\tau_p$, $\tau_d$ and $\tau_u$, respectively, should not exceed the coherence time/bandwidth of the channel, denoted as $\tau_c$, that is, $\tau_p+\tau_d+\tau_u \leq \tau_c$, with all these intervals specified in samples (or channel uses) on a time-frequency grid.

### A. CHANNEL MODEL

The propagation channel linking AP $m$ to MS $k$ is denoted by $g_{mk}$ and modelled as

$$g_{mk} = \sqrt{\beta_{mk}} h_{mk}, \tag{1}$$

where $\beta_{mk}$ represents the large-scale propagation losses (i.e., path loss and shadowing) and $h_{mk}$ corresponds to small-scale fading. The large-scale gain is further decomposed as $\beta_{mk} = \zeta_{mk}\chi_{mk}$ with $\zeta_{mk}$ representing the distance-dependent path loss and $\chi_{mk}$ corresponding to the shadowing component. Finally, the small-scale fading terms $h_{mk}$ consist of independent and identically distributed (i.i.d.) complex Gaussian random variables distributed as $\mathcal{CN}(0,1)$.

The channel coefficients $g_{mk}$ are assumed to be static throughout the coherence interval and then change independently (i.e., block fading). As in the seminal papers [4], [12] introducing the idea of cell-free operation, it is assumed that the CPU has perfect knowledge of the large-scale fading gains (i.e., $\beta_{mk}\ \forall mk$).

### B. A LINEAR MODEL FOR THE QUANTIZATION PROCESS

One of the major difficulties to be overcome when analyzing communication systems using low-resolution ADCs is the nonlinear nature of the scalar quantizers used to compress the signals to be transferred between APs and CPU during the different transmission phases. Let us denote by $\mathcal{Q}_\theta(\boldsymbol{y})$ the mathematical operations performed by a generic scalar quantizer $\theta$ on a generic signal vector $\boldsymbol{y}$. One of the classical approaches to deal with the nonlinear nature of $\mathcal{Q}_\theta(\boldsymbol{y})$ is the use of the Bussgang decomposition [31]

$$\hat{\boldsymbol{y}} = \mathcal{Q}_\theta(\boldsymbol{y}) = \boldsymbol{F}_\theta \boldsymbol{y} + \boldsymbol{q}_\theta, \tag{2}$$

where the matrix $\boldsymbol{F}_\theta$ can be obtained from the linear MMSE estimation of $\hat{\boldsymbol{y}}$ given $\boldsymbol{y}$, that is,

$$\boldsymbol{F}_\theta = \mathbb{E}\left\{\hat{\boldsymbol{y}}\boldsymbol{y}^H\right\}\left(\mathbb{E}\left\{\boldsymbol{y}\boldsymbol{y}^H\right\}\right)^{-1} = \boldsymbol{R}_{\hat{\boldsymbol{y}}\boldsymbol{y}}\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}^{-1}, \tag{3}$$

and $\boldsymbol{q}_\theta$ is a zero-mean additive quantization noise, uncorrelated with $\boldsymbol{y}$, and with correlation matrix

$$\begin{aligned}\boldsymbol{R}_{\boldsymbol{q}_\theta\boldsymbol{q}_\theta} &= \mathbb{E}\left\{\left(\hat{\boldsymbol{y}} - \boldsymbol{F}_\theta\boldsymbol{y}\right)\left(\hat{\boldsymbol{y}} - \boldsymbol{F}_\theta\boldsymbol{y}\right)^H\right\} \\ &= \boldsymbol{R}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}} - \boldsymbol{R}_{\hat{\boldsymbol{y}}\boldsymbol{y}}\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}^{-1}\boldsymbol{R}_{\boldsymbol{y}\hat{\boldsymbol{y}}}.\end{aligned} \tag{4}$$

The main challenge in using this decomposition consists of deriving the covariance matrices $\boldsymbol{R}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}}$ and $\boldsymbol{R}_{\hat{\boldsymbol{y}}\boldsymbol{y}}$ assuming that the input signal $\boldsymbol{y}$ is Gaussian with known covariance matrix $\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}$. For the 1-bit quantizer, these matrices can be found in a closed form [30, Section V]. For general scalar quantizers, however, these matrices are usually evaluated by resorting to approximations. In [30, Section IV], Mezghani and Nossek describe an approximation that is based on the use of the distortion factor $\rho_\theta = (1-\alpha_\theta) = 1/\,\mathrm{SQNR}_\theta$, where $\mathrm{SQNR}_\theta$ denotes the signal-to-quantization noise ratio. Note that the parameter $\alpha_\theta$ has only been introduced for notational convenience. In particular, they use a Gaussian approximation of $\boldsymbol{q}_\theta$ and show that (see [30, eqs. (25) and (28)])

$$\boldsymbol{R}_{\hat{\boldsymbol{y}}\boldsymbol{y}} = \boldsymbol{R}_{\boldsymbol{y}\hat{\boldsymbol{y}}} = \alpha_\theta \boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}, \tag{5}$$

and

$$\boldsymbol{R}_{\hat{\boldsymbol{y}}\hat{\boldsymbol{y}}} \approx \alpha_\theta^2 \boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}} + \alpha_\theta(1-\alpha_\theta)\,\mathrm{diag}\left(\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}\right), \tag{6}$$

with $\mathrm{diag}(\boldsymbol{X})$ being a diagonal matrix containing the diagonal of the square matrix $\boldsymbol{X}$ in its main diagonal. Hence, in summary, using (5) and (6) in (2), (3) and (4), the approach proposed by Mezghani and Nossek in [30], which is based on the Bussgang decomposition, allows us to approximate the vector of scalar quantized samples as

$$\hat{\boldsymbol{y}} = \mathcal{Q}_\theta(\boldsymbol{y}) \approx \alpha_\theta \boldsymbol{y} + \tilde{\boldsymbol{q}}_\theta, \tag{7}$$

where $\tilde{\boldsymbol{q}}_\theta \sim \mathcal{CN}\left(\boldsymbol{0}, \boldsymbol{R}_{\tilde{\boldsymbol{q}}_\theta\tilde{\boldsymbol{q}}_\theta}\right)$ with

$$\boldsymbol{R}_{\tilde{\boldsymbol{q}}_\theta\tilde{\boldsymbol{q}}_\theta} = \alpha_\theta(1-\alpha_\theta)\,\mathrm{diag}\left(\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}\right). \tag{8}$$

The optimal design parameters of uniform as well as non-uniform quantizers and the resulting distortion factor $\rho_\theta$ (or, equivalently, $\alpha_\theta$) are tabulated in [34] assuming Gaussian distributed input signals and different numbers of quantization bits per sample $b_\theta$. In particular, $\alpha_\theta$ is an increasing function of $b_\theta$ and, for the non-uniform quantizer case, which will be considered in the next sections, can be approximately expressed as $\alpha_\theta = 1 - \frac{\pi\sqrt{3}}{2}2^{-2\,b_\theta}$ for $b_\theta > 5$ [35], and the corresponding values for $b_\theta = 1, 2, 3, 4$ and 5 are summarized in Table 1 (derived from [34, Table 1]). This linear model for the quantization process (or its equivalent AQNM) has been extensively used in the massive MIMO and cell-free massive MIMO literature (see, among many others, [23], [25], [36]–[45] and references therein).

**TABLE 1. Parameters of the additive quantization noise model (derived from [34, Table 1]).**

| $b_\theta$ | $N_\theta = 2^{b_\theta}$ | $\rho_\theta$ | $\alpha_\theta = 1-\rho_\theta$ |
|---|---|---|---|
| 1 | 2 | 0.3634 | 0.6366 |
| 2 | 4 | 0.1175 | 0.8825 |
| 3 | 8 | 0.03454 | 0.96546 |
| 4 | 16 | 0.009497 | 0.990503 |
| 5 | 32 | 0.002499 | 0.997501 |

### C. TRAINING PHASE

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 either at the CPU, in the BCU-based case (see Fig. 1(a)), or at the APs, in the BAP-based approach (see Fig. 1(b)). During the UL training phase, all $K$ MSs simultaneously transmit pilot sequences of $\tau_p$ samples to the APs and thus, the $\tau_p \times 1$ received UL signal at the $m$th AP is given by

$$\boldsymbol{y}_{p_m} = \sqrt{\tau_p P_p}\sum_{k=1}^{K} g_{mk}\boldsymbol{\varphi}_k + \boldsymbol{n}_{p_m}, \tag{9}$$

where $P_p$ is the transmit power of each pilot symbol, $\boldsymbol{\varphi}_k$, with $\|\boldsymbol{\varphi}_k\|^2 = 1$, denotes the $\tau_p \times 1$ training sequence assigned to MS $k$, and $\boldsymbol{n}_{p_m}$ is a $\tau_p \times 1$ vector of i.i.d. additive noise samples with each entry distributed as $\mathcal{CN}(0, \sigma_u^2)$. For later use, note that the correlation matrix of $\boldsymbol{y}_{p_m}$ can be obtained as

$$\begin{aligned}\boldsymbol{R}_{\boldsymbol{y}_{p_m}\boldsymbol{y}_{p_m}} &= \mathbb{E}\left\{\boldsymbol{y}_{p_m}\boldsymbol{y}_{p_m}^H\right\} \\ &= \tau_p P_p \sum_{k=1}^{K}\beta_{mk}\boldsymbol{\varphi}_k\boldsymbol{\varphi}_k^H + \sigma_u^2\boldsymbol{I}_{\tau_p}.\end{aligned} \tag{10}$$

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 effect, a widely studied phenomenon in the context of centralized massive MIMO systems [46].

In the BCU-based approach, the components of vector $\boldsymbol{y}_{p_m}$ are compressed using a scalar quantizer at the AP and sent to the CPU, via the UL fronthaul link. Using the linear quantization noise model introduced in (2), the quantized signal vector at the CPU during the training phase can be expressed as

$$\begin{aligned}\hat{\boldsymbol{y}}_{pm} &= \mathcal{Q}_{pm}\left(\boldsymbol{y}_{p_m}\right)\\ &\approx \alpha_{p_m}\left(\sqrt{\tau_p P_p}\sum_{k=1}^{K} g_{mk}\boldsymbol{\varphi}_k + \boldsymbol{n}_{p_m}\right) + \tilde{\boldsymbol{q}}_{pm},\end{aligned} \tag{11}$$

where

$$\begin{aligned}\boldsymbol{R}_{\tilde{q}_{pm}\tilde{q}_{pm}} &= \mathbb{E}\left\{\tilde{\boldsymbol{q}}_{pm}\tilde{\boldsymbol{q}}_{pm}^H\right\}\\ &= \alpha_{p_m}\left(1-\alpha_{p_m}\right)\operatorname{diag}\left(\boldsymbol{R}_{\boldsymbol{y}_{p_m}\boldsymbol{y}_{p_m}}\right).\end{aligned} \tag{12}$$

### D. CHANNEL ESTIMATION

In order to estimate the channel for MS $k$, the signal vectors $\hat{\boldsymbol{y}}_{pm}$, in the BCU-based approach, or $\boldsymbol{y}_{p_m}$, in the BAP-based approach, are first projected onto the pilot signal $\boldsymbol{\varphi}_k^H$ to obtain

$$\check{y}_{p_{mk}} = \begin{cases}\boldsymbol{\varphi}_k^H\hat{\boldsymbol{y}}_{pm}, & \text{BCU}\\ \boldsymbol{\varphi}_k^H\boldsymbol{y}_{p_m}, & \text{BAP.}\end{cases} \tag{13}$$

Given $\check{y}_{p_{mk}}$, the linear MMSE estimate of $g_{mk}$ can then be calculated as[2] [4]

$$\begin{aligned}\hat{g}_{mk} &= \frac{\mathbb{E}\left\{\check{y}_{p_{mk}}^* g_{mk}\right\}}{\mathbb{E}\left\{\left|\check{y}_{p_{mk}}\right|^2\right\}}\check{y}_{p_{mk}}\\ &= \begin{cases}\dfrac{\alpha_{p_m}\sqrt{\tau_p P_p}\beta_{mk}}{\alpha_{p_m}\xi_{mk}}\boldsymbol{\varphi}_k^H\hat{\boldsymbol{y}}_{pm}, & \text{BCU}\\ \dfrac{\sqrt{\tau_p P_p}\beta_{mk}}{\xi_{mk}}\boldsymbol{\varphi}_k^H\boldsymbol{y}_{p_m}, & \text{BAP,}\end{cases}\end{aligned} \tag{14}$$

where

$$\xi_{mk} = \tau_p P_p\sum_{k'=1}^{K}\beta_{mk'}\left|\boldsymbol{\varphi}_k^H\boldsymbol{\varphi}_{k'}\right|^2 + \sigma_u^2, \tag{15}$$

and we have used the fact that

$$\begin{aligned}\mathbb{E}\left\{\check{y}_{p_{mk}}^* g_{mk}\right\} &= \begin{cases}\mathbb{E}\left\{\left(\boldsymbol{\varphi}_k^H\hat{\boldsymbol{y}}_{pm}\right)^* g_{mk}\right\}, & \text{BCU}\\ \mathbb{E}\left\{\left(\boldsymbol{\varphi}_k^H\boldsymbol{y}_{p_m}\right)^* g_{mk}\right\}, & \text{BAP}\end{cases}\\ &= \begin{cases}\alpha_{p_m}\sqrt{\tau_p P_p}\beta_{mk}, & \text{BCU}\\ \sqrt{\tau_p P_p}\beta_{mk}, & \text{BAP,}\end{cases}\end{aligned} \tag{16}$$

and

$$\begin{aligned}\mathbb{E}\left\{\left|\check{y}_{p_{mk}}\right|^2\right\} &= \begin{cases}\boldsymbol{\varphi}_k^H\mathbb{E}\left\{\hat{\boldsymbol{y}}_{pm}\hat{\boldsymbol{y}}_{pm}^H\right\}\boldsymbol{\varphi}_k, & \text{BCU}\\ \boldsymbol{\varphi}_k^H\mathbb{E}\left\{\boldsymbol{y}_{p_m}\boldsymbol{y}_{p_m}^H\right\}\boldsymbol{\varphi}_k, & \text{BAP}\end{cases}\\ &= \begin{cases}\boldsymbol{\varphi}_k^H\left(\alpha_{p_m}^2\boldsymbol{R}_{\boldsymbol{y}_{p_m}\boldsymbol{y}_{p_m}} + \boldsymbol{R}_{\tilde{q}_{pm}\tilde{q}_{pm}}\right)\boldsymbol{\varphi}_k, & \text{BCU}\\ \boldsymbol{\varphi}_k^H\boldsymbol{R}_{\boldsymbol{y}_{p_m}\boldsymbol{y}_{p_m}}\boldsymbol{\varphi}_k, & \text{BAP}\end{cases}\\ &= \begin{cases}\alpha_{p_m}\xi_{mk}, & \text{BCU}\\ \xi_{mk}, & \text{BAP,}\end{cases}\end{aligned} \tag{17}$$

Moreover, note that $g_{mk} = \hat{g}_{mk} + \tilde{g}_{mk}$ where the channel estimate $\hat{g}_{mk}$ and the channel estimation error $\tilde{g}_{mk}$ are mutually independent.

### E. UPLINK PAYLOAD DATA TRANSMISSION

The signal transmitted by the $k$th MS is $z_k = \sqrt{\nu_k}s_{uk}$, with $0 \le \nu_k \le P_{uk}$ and $\mathbb{E}\left\{|s_{uk}|^2\right\} = 1$, where $s_{uk}$ is the transmitted symbol, $\nu_k$ denotes the transmitted power, and $P_{uk}$ is the maximum average transmitted power available at the $k$th MS. Using this notation, the received signal at the $m$th AP can be expressed as

$$r_{um} = \sum_{k'=1}^{K}\sqrt{\nu_{k'}}g_{mk'}s_{uk'} + n_{um}, \tag{18}$$

where $n_{um} \sim \mathcal{CN}(0, \sigma_u^2)$ is the additive thermal noise sample at the receiver output. The received signal at each of the APs in the network is subject to signal processing operations that depend on the particular CPU-AP functional split under consideration. The combined signal at the CPU corresponding to the symbol transmitted by the $k$th MS can be written as

$$y_{uk} = \sum_{m=1}^{M} x_{umk}, \tag{19}$$

where (see Figs. 1(c) and 1(d))

$$\begin{aligned}x_{umk} &= \begin{cases}\varpi_{umk}\mathcal{Q}_{um}\left(r_{um}\right), & \text{BCU}\\ \mathcal{Q}_{umk}\left(\varpi_{umk}r_{um}\right), & \text{BAP,}\end{cases}\\ &\approx \begin{cases}\varpi_{umk}\left(\alpha_{um}r_{um} + \tilde{q}_{um}\right), & \text{BCU}\\ \alpha_{umk}\varpi_{umk}r_{um} + \tilde{q}_{umk}, & \text{BAP,}\end{cases}\end{aligned} \tag{20}$$

with $\varpi_{umk}$ denoting the decoding coefficient applied by the $m$th AP to the signal from MS $k$. Note that the Bussgang decomposition assumes that the input signal to the quantizer has a Gaussian distribution. Since the input to the quantizer is equal to the sum of many random variables, the central limit theorem ensures that it has a near Gaussian distribution thus justifying the use of this decomposition. The Gaussian approximation was numerically verified by Bashar *et al.*, for typical parameter values, as shown in [23, Figs. 2(a)-2(c)]. Now, assuming, as in [4], that the CPU uses only statistical CSI when detecting the transmitted symbol $s_{uk}$ (this is a well-known strategy in the context of massive MIMO performance analysis that is denoted as the *use and then forget*

[2]Note that the signal at the output of a quantizer is not Gaussian and thus, in the specific case of BCU where the channel estimation process deals with quantized samples, a linear MMSE is no longer an MMSE estimate. Following the additive Gaussian quantization noise model proposed by Mezghani and Nossek in [30], which largely relies on the Gaussian approximation, we adopt a conservative approach and assume that the estimation and estimate behave as uncorrelated Gaussian, and thus independent, random vectors. Although this approximation does not completely supersede the need for an optimal MMSE estimator, results in [23], [30], [47] demonstrate the effectiveness of this approach.

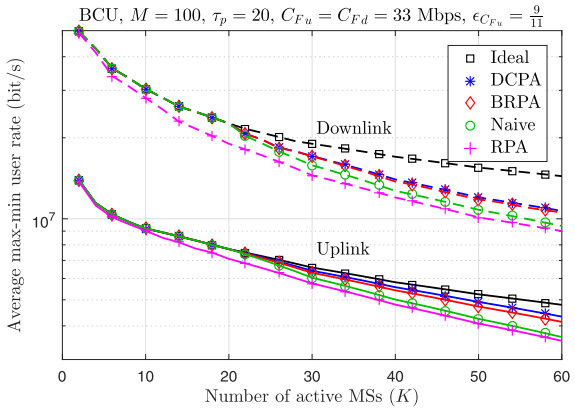



**FIGURE 2.** Average max-min rate per-user versus the number of active MSs under different pilot allocation strategies (BCU, $M = 100$ APs, $\tau_p = 20$ samples, $C_{Fu} = C_{Fd} = 33$ Mbps, $\epsilon_{C_{Fu}} = 9/11$).

*CSI* scheme (see, for instance, [2, Section 3.2.1]), the combined signal can be rewritten as

$$y_{uk} = \mathrm{DS}_{uk} s_{uk} + \mathrm{BU}_{uk} s_{uk} + \sum_{k' \neq k} \mathrm{UI}_{ukk'} s_{uk'} + \mathrm{QN}_{uk} + w_{uk}, \quad (21)$$

where, for notational convenience, we define

$$\tilde{\alpha}_{umk} = \begin{cases} \alpha_{um}, & \text{BCU}, \\ \alpha_{umk}, & \text{BAP}, \end{cases} \quad (22)$$

for all $k$, allowing us to express the strength of the desired signal (DS), the beamforming gain uncertainty (BU), the interference caused by the $k'$th user (UI), the quantization noise (QN) due to the use of capacity-constrained fronthaul links, and the combined additive white Gaussian noise as

$$\mathrm{DS}_{uk} = \sqrt{\nu_k} \sum_{m=1}^{M} \tilde{\alpha}_{umk} \mathbb{E}\{\varpi_{umk} g_{mk}\}, \quad (23)$$

$$\mathrm{BU}_{uk} = \sqrt{\nu_k} \sum_{m=1}^{M} \tilde{\alpha}_{umk} \left(\varpi_{umk} g_{mk} - \mathbb{E}\{\varpi_{umk} g_{mk}\}\right), \quad (24)$$

$$\mathrm{UI}_{ukk'} = \sqrt{\nu_{k'}} \sum_{m=1}^{M} \tilde{\alpha}_{umk} \varpi_{umk} g_{mk'}, \quad (25)$$

$$\mathrm{QN}_{uk} = \begin{cases} \sum_{m=1}^{M} \varpi_{umk} \tilde{q}_{um}, & \text{BCU} \\ \sum_{m=1}^{M} \tilde{q}_{umk}, & \text{BAP}, \end{cases} \quad (26)$$

and

$$w_{uk} = \sum_{m=1}^{M} \tilde{\alpha}_{umk} \varpi_{umk} n_{um}, \quad (27)$$

respectively.

### F. DOWNLINK PAYLOAD DATA TRANSMISSION

Let us define by $\boldsymbol{s}_d = [s_{d1} \ldots s_{dK}]^T$ the $K \times 1$ vector of symbols jointly (cooperatively) transmitted from the APs to the MSs, such that $\mathbb{E}\{\boldsymbol{s}_d \boldsymbol{s}_d^H\} = \boldsymbol{I}_K$. The signal processing that symbol vector $\boldsymbol{s}_d$ undergoes before being transmitted depends on the implemented CPU-AP functional split, as previously described in the introduction to this section. In particular, the mathematical operations performed to obtain the signal to be transmitted from the $m$th AP, denoted as $x_{dm}$, can be summarized as (see Figs. 1(e) and 1(f))

$$x_{dm} = \begin{cases} \mathcal{Q}_{dm}\left(\sum_{k=1}^{K} \eta_{mk}^{1/2} \varpi_{dmk} s_{dk}\right), & \text{BCU} \\ \sum_{k=1}^{K} \mathcal{Q}_{dmk}\left(\eta_{mk}^{1/2} s_{dk}\right) \varpi_{dmk}, & \text{BAP}, \end{cases}$$
$$\approx \begin{cases} \alpha_{dm} \sum_{k=1}^{K} \eta_{mk}^{1/2} \varpi_{dmk} s_{dk} + \tilde{q}_{dm}, & \text{BCU} \\ \sum_{k=1}^{K} \varpi_{dmk}\left(\alpha_{dmk} \eta_{mk}^{1/2} s_{dk} + \tilde{q}_{dmk}\right), & \text{BAP}, \end{cases} \quad (28)$$

where $\varpi_{dmk}$ is the beamforming coefficient for MS $k$ applied to the signal received by the $m$th AP, and $\eta_{mk}$ is the corresponding power allocation coefficient, chosen to meet a prescribed criterion while satisfying the power constraint $\mathbb{E}\{|x_{dm}|^2\} \leq P_{dm}$ at the $m$th AP, with $P_{dm}$ denoting the maximum average transmit power available at the $m$th AP. Hence, taking into account that

$$\mathbb{E}\{|\tilde{q}_{dm}|^2\} = \alpha_{dm}(1 - \alpha_{dm}) \mathbb{E}\left\{\left|\sum_{k=1}^{K} \eta_{mk}^{1/2} \varpi_{dmk} s_{dk}\right|^2\right\}$$
$$= \alpha_{dm}(1 - \alpha_{dm}) \sum_{k=1}^{K} \eta_{mk} \mathbb{E}\{|\varpi_{dmk}|^2\}, \quad (29)$$

$$\mathbb{E}\{|\tilde{q}_{dmk}|^2\} = \alpha_{dmk}(1 - \alpha_{dmk}) \mathbb{E}\left\{\left|\eta_{mk}^{1/2} s_{dk}\right|^2\right\}$$
$$= \alpha_{dmk}(1 - \alpha_{dmk}) \eta_{mk}, \quad (30)$$

we have that

$$\mathbb{E}\{|x_{dm}|^2\} = \begin{cases} \alpha_{dm}^2 \sum_{k'=1}^{K} \eta_{mk'} \mathbb{E}\{|\varpi_{dmk'}|^2\} + \mathbb{E}\{|\tilde{q}_{dm}|^2\}, & \text{BCU} \\ \sum_{k'=1}^{K} \left(\alpha_{dmk'}^2 \eta_{mk'} + \mathbb{E}\{|\tilde{q}_{dmk'}|^2\}\right) \mathbb{E}\{|\varpi_{dmk'}|^2\}, & \text{BAP} \end{cases}$$
$$= \begin{cases} \alpha_{dm} \sum_{k'=1}^{K} \eta_{mk'} \mathbb{E}\{|\varpi_{dmk'}|^2\}, & \text{BCU} \\ \sum_{k'=1}^{K} \alpha_{dmk'} \eta_{mk'} \mathbb{E}\{|\varpi_{dmk'}|^2\}, & \text{BAP}. \end{cases} \quad (31)$$

Hence, using the definition

$$\tilde{\alpha}_{dmk} = \begin{cases} \alpha_{dm}, & \text{BCU}, \\ \alpha_{dmk}, & \text{BAP}, \end{cases} \quad (32)$$

for all $k$, the DL power allocation constraints can be expressed for both BCU- and BAP-based functional splits as

$$\sum_{k=1}^{K} \tilde{\alpha}_{dmk} \eta_{mk} \mathbb{E}\{|\varpi_{dmk}|^2\} \leq P_{dm}. \quad (33)$$

The signal received by MS $k$ can be expressed as

$$y_{dk} = \sum_{m=1}^{M} g_{mk} x_{dm} + n_{dk}, \tag{34}$$

where $n_{dk} \sim \mathcal{CN}(0, \sigma_d^2)$ is the corresponding Gaussian noise sample. Under the assumption that only statistical CSI is available at the MSs, the received signal at the $k$th MS can be rewritten as [4]

$$y_{dk} = \mathrm{DS}_{dk} s_{dk} + \mathrm{BU}_{dk} s_{dk} + \sum_{k' \neq k} \mathrm{UI}_{dkk'} s_{dk'} + \mathrm{QN}_{dk} + w_{dk}, \tag{35}$$

where

$$\mathrm{DS}_{dk} = \sum_{m=1}^{M} \tilde{\alpha}_{dmk} \eta_{mk}^{1/2} \mathbb{E}\{g_{mk} \varpi_{dmk}\}, \tag{36}$$

$$\mathrm{BU}_{dk} = \sum_{m=1}^{M} \tilde{\alpha}_{dmk} \eta_{mk}^{1/2} (g_{mk} \varpi_{dmk} - \mathbb{E}\{g_{mk} \varpi_{dmk}\}), \tag{37}$$

$$\mathrm{UI}_{dkk'} = \sum_{m=1}^{M} \tilde{\alpha}_{dmk'} \eta_{mk'}^{1/2} g_{mk} \varpi_{dmk'}, \tag{38}$$

$$\mathrm{QN}_{dk} = \begin{cases} \sum_{m=1}^{M} g_{mk} \tilde{q}_{dm}, & \text{BCU} \\ \sum_{m=1}^{M} g_{mk} \sum_{k'=1}^{K} \varpi_{dmk'} \tilde{q}_{dmk'}, & \text{BAP}, \end{cases} \tag{39}$$

and $w_{dk} = n_{dk}$, represent the strength of the desired signal (DS), the beamforming gain uncertainty (BU), the interuser interference (UI) caused by the transmission to the $k'$th MS, the quantization noise (QN) and the thermal noise, respectively.

### G. PRECODING/DECODING SCHEMES

It is well-known that cellular massive MIMO systems using simple CB precoders in the DL and MF decoders in the UL are able to provide high spectral efficiency by relying on *channel hardening* and *favorable propagation* phenomena [2]. Under channel hardening conditions, there is no need to adapt the radio resource management functions (i.e., power control and scheduling) to the small-scale fading variations. Furthermore, favorable propagation conditions guarantee that propagation channels observed by different MSs are almost orthogonal and, thus, the presence of little inter-user interference leakage. In a cell-free massive MIMO network, however, it has been recently shown in [28] that neither channel hardening nor favourable propagation conditions can always be guaranteed using simple CB schemes with single-antenna APs. Consequently, one should not rely on these propagation assumptions when obtaining the achievable rates, as this could lead to a great underestimation of the achievable performance [28].

In [29], Interdonato *et al.* proposed a DL precoding scheme, named NCB, that satisfies short-term average power constraints at the APs and, most importantly, it largely improves the channel hardening and favourable propagation conditions when compared with the *classical* CB scheme. In the UL, in contrast, the classical MF solution outperforms the normalized counterpart. That is, using the NCB/MF precoding/decoding setup, defined as

$$\varpi_{dmk} = \frac{\hat{g}_{mk}^*}{|\hat{g}_{mk}|}, \tag{40a}$$

$$\varpi_{umk} = \hat{g}_{mk}^*, \tag{40b}$$

respectively, we can safely rely on channel hardening and favourable propagation conditions when obtaining the achievable rates without the risk of underestimating the achievable performance even when using single-antenna APs.

## III. ACHIEVABLE RATES

Analysis techniques similar to those applied, for instance, in [2], [4], [29], are used in this section to derive UL and DL achievable rates. In particular, the sum of the second, third, fourth and fifth terms in (21), for the UL case, or (35), for the DL case, is treated as *effective noise*. In fact, as the data symbols transmitted by different MSs are mutually uncorrelated and are also uncorrelated with both the quantization and thermal noise samples, it can be shown that the additive terms constituting the *effective noise* are, in both UL and DL cases, mutually uncorrelated, and uncorrelated with the desired signal term[3] [4], [23], [47]. Hence, the power of the effective noise may be treated as the sum of the powers of these terms. Now, recalling the fact that uncorrelated Gaussian noise represents the worst case [4], the UL and DL achievable rates (measured in bits per second) for MS $k$ can be obtained as in (41), as shown at the bottom of the next page, where $B$ is the bandwidth and $l$ is a token used to represent either the DL (with $l = d$) or the UL (with $l = u$).

Irrespective of whether we are using the UL MF approach or the DL non-cooperative NCB, the expectations and variances in (41) can be calculated in closed-form. For notational convenience, let us define the variable $\varrho_{mk}$ as

$$\varrho_{mk} \triangleq \mathbb{E}\left\{|\hat{g}_{mk}|^2\right\} = \begin{cases} \dfrac{\alpha_{p_m} \tau_p P_p \beta_{mk}^2}{\xi_{mk}}, & \text{BCU}, \\ \dfrac{\tau_p P_p \beta_{mk}^2}{\xi_{mk}}, & \text{BAP}. \end{cases} \tag{42}$$

[3]Note that all the terms in the received signal are zero-mean because, as already specified in the paper, $\mathbb{E}\{s_{lk}\} = \mathbb{E}\{\tilde{q}_{lmk}\} = \mathbb{E}\{w_{lk}\} = 0$, for all $l \in \{u, d\}$, $m \in \{1, \ldots, M\}$ and $k \in \{1, \ldots, K\}$. Furthermore, using a similar reasoning as that exploited by Ngo *et al.* in [4], since $s_{lk}$ is independent of $\mathrm{DS}_{lk}$ and $\mathrm{BU}_{lk}$, we have that $\mathbb{E}\{\mathrm{DS}_{lk} s_{lk} \mathrm{BU}_{lk}^* s_{lk}^*\} = \mathbb{E}\{\mathrm{DS}_{lk} \mathrm{BU}_{lk}^*\} \mathbb{E}\{|s_{lk}|^2\} = 0$. Thus, the first and second terms of the received signal are uncorrelated. An analogous calculation shows that $\mathbb{E}\{\mathrm{DS}_{lk} s_{lk} \mathrm{UI}_{lkk'}^* s_{lk'}^*\} = \mathbb{E}\{\mathrm{DS}_{lk} \mathrm{UI}_{lkk'}^*\} \mathbb{E}\{s_{lk} s_{lk'}^*\} = 0$, for all $k \neq k'$, thus proving that the first and third terms in the received signal are also uncorrelated. In general, using the fact that $\mathbb{E}\{s_{lk} s_{lk'}^*\} = 0$, for all $k' \neq k$, $\mathbb{E}\{s_{lk'} \tilde{q}_{lk}^*\} = \mathbb{E}\{s_{lk'} w_{lk}^*\} = 0$, for all $k$ and $k'$, it can easily be shown that all terms in the received signal are uncorrelated.

### A. CALCULATION OF $|\mathrm{DS}_{lk}|^2$

As shown in Appendix A, the term $|\mathrm{DS}_{lk}|^2$ in the numerator of (41) can be calculated in an exact closed-form as

$$|\mathrm{DS}_{uk}|^2 = \nu_k \left( \sum_{m=1}^{M} \tilde{\alpha}_{umk} \mathbb{E}\{g_{mk} \varpi_{umk}\} \right)^2 = \nu_k \left( \sum_{m=1}^{M} \tilde{\alpha}_{umk} \varrho_{mk} \right)^2, \quad (43)$$

for the UL case, and as

$$|\mathrm{DS}_{dk}|^2 = \left( \sum_{m=1}^{M} \tilde{\alpha}_{dmk} \eta_{mk}^{1/2} \mathbb{E}\{g_{mk} \varpi_{dmk}\} \right)^2 = \frac{\pi}{4} \left( \sum_{m=1}^{M} \tilde{\alpha}_{dmk} \sqrt{\eta_{mk} \varrho_{mk}} \right)^2, \quad (44)$$

for the DL case.

### B. CALCULATION OF $\mathbb{E}\left\{|\mathrm{BU}_{lk}|^2\right\}$

Again, it is shown in Appendix A that the term $\mathbb{E}\left\{|\mathrm{BU}_{lk}|^2\right\}$ in the denominator of (41) can be calculated in an exact closed-form as

$$\mathbb{E}\left\{|\mathrm{BU}_{uk}|^2\right\} = \nu_k \sum_{m=1}^{M} \tilde{\alpha}_{umk}^2 \operatorname{Var}\{g_{mk} \varpi_{umk}\} = \nu_k \sum_{m=1}^{M} \tilde{\alpha}_{umk}^2 \varrho_{mk} \beta_{mk}, \quad (45)$$

for the UL case, and

$$\mathbb{E}\left\{|\mathrm{BU}_{dk}|^2\right\} = \sum_{m=1}^{M} \tilde{\alpha}_{dmk}^2 \eta_{mk} \operatorname{Var}\{g_{mk} \varpi_{dmk}\} = \sum_{m=1}^{M} \tilde{\alpha}_{dmk}^2 \eta_{mk} \left( \beta_{mk} - \frac{\pi}{4} \varrho_{mk} \right), \quad (46)$$

for the DL case.

### C. CALCULATION OF $\mathbb{E}\left\{|\mathrm{UI}_{lkk'}|^2\right\}$

In Appendix A, it is also shown that the statistical expectations $\mathbb{E}\left\{|\mathrm{UI}_{lkk'}|^2\right\}$ in the denominator of (41) can be calculated as

$$\mathbb{E}\left\{|\mathrm{UI}_{ukk'}|^2\right\} = \nu_{k'} \sum_{m=1}^{M} \sum_{n=1}^{M} \tilde{\alpha}_{umk} \tilde{\alpha}_{unk} \mathbb{E}\left\{g_{mk'} g_{nk'}^* \varpi_{umk} \varpi_{unk}^*\right\} = \nu_{k'} \sum_{m=1}^{M} \tilde{\alpha}_{umk}^2 \varrho_{mk} \beta_{mk'} + \nu_{k'} \left( \sum_{m=1}^{M} \tilde{\alpha}_{umk} \frac{\beta_{mk}}{\beta_{mk'}} \varrho_{mk'} \right)^2 \left| \boldsymbol{\varphi}_{k'}^H \boldsymbol{\varphi}_k \right|^2, \quad (47)$$

for the UL case, and

$$\mathbb{E}\left\{|\mathrm{UI}_{dkk'}|^2\right\} = \sum_{m=1}^{M} \sum_{n=1}^{M} \tilde{\alpha}_{dmk'} \tilde{\alpha}_{unk'} \eta_{mk'}^{1/2} \eta_{nk'}^{1/2} \mathbb{E}\left\{g_{mk} g_{nk}^* \varpi_{dmk'} \varpi_{dnk'}^*\right\} = \sum_{m=1}^{M} \tilde{\alpha}_{dmk'}^2 \eta_{mk'} \beta_{mk} + \frac{\pi}{4} \sum_{m=1}^{M} \sum_{\substack{n=1 \\ n \neq m}}^{M} \tilde{\alpha}_{dmk'} \tilde{\alpha}_{dnk'} \sqrt{\eta_{mk'} \eta_{nk'} \varrho_{mk} \varrho_{nk}} \left| \boldsymbol{\varphi}_{k'}^H \boldsymbol{\varphi}_k \right|^2, \quad (48)$$

for the DL case. Note that, in contrast to what was done by Interdonato *et al.* in [29, eq. (16)], instead of using approximations based on the first-order Taylor expansion of a quotient, we provide an exact closed-form expression for the inter-user interference term.

### D. CALCULATION OF $\mathbb{E}\left\{|\mathrm{QN}_{lk}|^2\right\}$

The quantization noise term also depends on the implemented CPU-AP functional split and, as shown in Appendix B, it can be approximated in closed form as

$$\mathbb{E}\left\{|\mathrm{QN}_{uk}|^2\right\} \simeq \sum_{m=1}^{M} \tilde{\alpha}_{umk} \left(1 - \tilde{\alpha}_{umk}\right) \left( \sum_{k'=1}^{K} \nu_{k'} \beta_{mk'} + \sigma_u^2 \right), \quad (49)$$

for the UL BCU-based case, as

$$\mathbb{E}\left\{|\mathrm{QN}_{uk}|^2\right\} \simeq \sum_{m=1}^{M} \tilde{\alpha}_{umk} \left(1 - \tilde{\alpha}_{umk}\right) \left( \sum_{k'=1}^{K} \nu_{k'} \beta_{mk'} \varrho_{mk} + \sum_{k'=1}^{K} \nu_{k'} \frac{\beta_{mk}^2}{\beta_{mk'}^2} \varrho_{mk'}^2 \left| \boldsymbol{\varphi}_{k'}^H \boldsymbol{\varphi}_k \right|^2 + \varrho_{mk} \sigma_u^2 \right), \quad (50)$$

for the UL BAP-based case, and as

$$\mathbb{E}\left\{|\mathrm{QN}_{dk}|^2\right\} \simeq \sum_{m=1}^{M} \beta_{mk} \sum_{k'=1}^{K} \tilde{\alpha}_{dmk'} \left(1 - \tilde{\alpha}_{dmk'}\right) \eta_{mk'}, \quad (51)$$

for the DL case.

$$R_{lk} = B \frac{\tau_l}{\tau_c} \log_2 \left( 1 + \frac{|\mathrm{DS}_{lk}|^2}{\mathbb{E}\left\{|\mathrm{BU}_{lk}|^2\right\} + \sum_{k' \neq k} \mathbb{E}\left\{|\mathrm{UI}_{lkk'}|^2\right\} + \mathbb{E}\left\{|\mathrm{QN}_{lk}|^2\right\} + \mathbb{E}\left\{|w_{lk}|^2\right\}} \right) \quad (41)$$

### E. CALCULATION OF $\mathbb{E}\left\{|w_{lk}|^2\right\}$

In the UL scenario we have that

$$\mathbb{E}\left\{|w_{uk}|^2\right\} = \sum_{m=1}^{M} \tilde{\alpha}_{umk}^2 \varrho_{mk} \sigma_u^2. \tag{52}$$

In the DL case, in contrast, $\mathbb{E}\left\{|w_{dk}|^2\right\} = \sigma_d^2$.

## IV. FRONTHAUL BANDWIDTH CONSUMPTION

Using the additive quantization noise model described in Subsect. II-B, let us denote by $b_\theta = b(\alpha_\theta)$ the number of bits used by a generic scalar quantizer $\mathcal{Q}_\theta$ to represent each of the input samples. Using this notation, we explicitly relate the number of quantization bits per sample with the parameter $\alpha_\theta$ characterizing the accuracy of the ADC.

### A. BCU-BASED CPU-AP FUNCTIONAL SPLIT

In the UL of a BCU-based CPU-AP functional split there are two sets of compressed data that must be conveyed from the APs to the CPU through the UL fronthaul link, namely, the received signal samples during the UL payload data transmission phase and the received signal vector during the training phase. During the UL payload data transmission phase, the compressed/decompressed signal sample per channel use on the $m$th UL fronthaul link is given by $\hat{x}_{um} = \mathcal{Q}_{um}(r_{um})$, and the required average rate (in bit/s) to transfer these compressed signals on the $m$th UL fronthaul link can then be obtained as

$$\hat{C}_{um}^{\text{Payload}} = B\frac{\tau_u}{\tau_c} b(\alpha_{um}) = B\frac{\tau_u}{\tau_c} b(\tilde{\alpha}_{umk}), \tag{53}$$

where the term $\tau_u/\tau_c$ accounts for the fact that only $\tau_u$ out of $\tau_c$ channel uses are employed for UL transmission. The unquantized signal vector at the CPU corresponding to the signal received by the $m$th AP during the training phase is given by (11). As this vector contains $\tau_p$ samples, the required average rate (in bit/s) to transfer the quantized vector $\hat{\boldsymbol{y}}_{pm} = \mathcal{Q}_{pm}\left(\boldsymbol{y}_{pm}\right)$ on the $m$th UL fronthaul link is given by

$$\hat{C}_{um}^{\text{Training}} = B\frac{\tau_p}{\tau_c} b\left(\alpha_{p_m}\right). \tag{54}$$

The only quantized signal per channel use to be transferred on the $m$th DL fronthaul link is $\hat{x}_{dm} = \mathcal{Q}_{dm}\left(\sum_{k=1}^{K} \eta_{mk}^{1/2} \varpi_{dmk} s_{dk}\right)$. Hence, the required average rate (in bit/s) to transfer the $\tau_d$ quantized signals on the $m$th DL fronthaul link is

$$\hat{C}_{dm} = B\frac{\tau_d}{\tau_c} b(\alpha_{dm}) = B\frac{\tau_d}{\tau_c} b(\tilde{\alpha}_{dmk}). \tag{55}$$

### B. BAP-BASED CPU-AP FUNCTIONAL SPLIT

In the UL of the BAP-based CPU-AP functional split, the quantized signal corresponding to the $k$th active MS can be expressed as $\hat{x}_{umk} = \mathcal{Q}_{umk}(\varpi_{umk} r_{um}) = \mathcal{Q}_{umk}\left(\hat{g}_{mk}^* r_{um}\right)$. The required average rate (measured in bit/s) to transfer each of these quantized signals on the $m$th UL fronthaul link is

$$\hat{C}_{umk} = B\frac{\tau_u}{\tau_c} b(\alpha_{umk}) = B\frac{\tau_u}{\tau_c} b(\tilde{\alpha}_{umk}), \tag{56}$$

for all $k$. In the DL of a BAP-based approach, the quantized signal corresponding to the $k$th active MS is $\hat{x}_{dmk} = \mathcal{Q}_{dmk}\left(\eta_{mk}^{1/2} s_{dk}\right)$. The required average rate (in bit/s) to transfer each of these compressed signals on the $m$th DL fronthaul link is

$$\hat{C}_{dmk} = B\frac{\tau_d}{\tau_c} b(\alpha_{dmk}) = B\frac{\tau_d}{\tau_c} b(\tilde{\alpha}_{dmk}). \tag{57}$$

## V. MAX-MIN POWER ALLOCATION AND OPTIMAL QUANTIZATION

### A. UPLINK POWER CONTROL AND QUANTIZATION

Max-min UL power allocation and quantization problems aim at finding the vector of power control coefficients $\boldsymbol{\nu} = [\nu_1 \ \ldots \ \nu_K]^T$, the vector of payload quantization accuracy parameters $\tilde{\boldsymbol{\alpha}}_u = [\tilde{\boldsymbol{\alpha}}_{u1} \ \ldots \ \tilde{\boldsymbol{\alpha}}_{uM}]^T$ and, for the BCU case, the vector of channel estimation quantization accuracy parameters $\boldsymbol{\alpha}_p = [\alpha_{p1} \ \ldots \ \alpha_{pM}]^T$, that jointly maximize the minimum of the achievable UL rates of all MSs while satisfying the transmit power constraints at each MS and the UL fronthaul bandwidth constraints at each AP [4], [12], [14], [23], [27].

#### 1) BAP-BASED CASE

Maximizing the minimum achievable rate per-user is equivalent to maximizing the minimum achievable signal-to-interference-plus-noise ratio (SINR) per-user. Consequently, the UL power control and quantization optimization problem can be formulated, for the BAP-based case, as

$$\begin{aligned} &\max_{(\boldsymbol{\nu},\, \tilde{\boldsymbol{\alpha}}_u)} \ \min_k \ \text{SINR}(\boldsymbol{\nu},\, \tilde{\boldsymbol{\alpha}}_u), \\ &\text{subject to } 0 \le \nu_k \le P_{uk} \quad \forall k, \\ &\qquad B\frac{\tau_u}{\tau_c} b(\alpha_{um}) \le \frac{C_{Fu}}{K} \quad \forall m, \end{aligned} \tag{58}$$

where

$$\text{SINR}(\boldsymbol{\nu},\, \tilde{\boldsymbol{\alpha}}_u) = \frac{\nu_k A_{mk}(\tilde{\boldsymbol{\alpha}}_u)}{\sum_{k'=1}^{K} \nu_{k'} B_{kk'}(\tilde{\boldsymbol{\alpha}}_u) + \sum_{k' \ne k} \nu_{k'} C_{kk'}(\tilde{\boldsymbol{\alpha}}_u) + D_k(\tilde{\boldsymbol{\alpha}}_u)} \tag{59}$$

with

$$A_{mk}(\tilde{\boldsymbol{\alpha}}_u) = \left(\sum_{m=1}^{M} \tilde{\alpha}_{umk} \varrho_{mk}\right)^2, \tag{60}$$

$$\begin{aligned} B_{kk'}(\tilde{\boldsymbol{\alpha}}_u) = &\sum_{m=1}^{M} \tilde{\alpha}_{umk} \beta_{mk'} \varrho_{mk} \\ &+ \sum_{m=1}^{M} \tilde{\alpha}_{umk}(1 - \tilde{\alpha}_{umk}) \frac{\beta_{mk}^2}{\beta_{mk'}^2} \varrho_{mk'}^2 \left|\boldsymbol{\varphi}_{k'}^H \boldsymbol{\varphi}_k\right|^2, \end{aligned} \tag{61}$$

$$C_{kk'}(\tilde{\boldsymbol{\alpha}}_u) = \left(\sum_{m=1}^{M} \tilde{\alpha}_{umk} \frac{\beta_{mk}}{\beta_{mk'}} \varrho_{mk'}\right)^2 \left|\boldsymbol{\varphi}_{k'}^H \boldsymbol{\varphi}_k\right|^2, \tag{62}$$

and

$$D_k(\tilde{\boldsymbol{\alpha}}_u) = \sum_{m=1}^{M} \tilde{\alpha}_{umk} \varrho_{mk} \sigma_u^2. \tag{63}$$

The objective function is the achievable SINR of MS $k$ as calculated in Section III (see, (43), (45), (47), (50), and (52)). The first constraint indicates that the power allocated to MS $k$ must be less than or equal to the available transmit power. The second constraint implies that the average rate used to transfer the quantized signal of user $k$ on the $m$th fronthaul link, as calculated in (56), must not be higher than $C_{Fu}/K$. That is, it is assumed that the signals from different MSs are quantized with the same amount of bit/sample (i.e., the available UL fronthaul capacity $C_{Fu}$ is equally split among the $K$ active MSs at all APs). Although there are other strategies that could be implemented in practice, the problem of devising an optimal fronthaul capacity allocation approach is left for further research.

As the achievable rates monotonically increase with increasing quantizer quality, under optimal conditions the number of quantization bits $b(\alpha_{um})$ will be as large as possible given the fronthaul link capacity constraint, that is,

$$b\left(\alpha_{um}^{\text{opt}}\right) = \left\lfloor \frac{C_{Fu}\tau_c}{K\tau_u} \right\rfloor \tag{64}$$

or, equivalently,

$$\tilde{\alpha}_{umk}^{\text{opt}} = \alpha_{um}^{\text{opt}} = b^{-1}\left(\left\lfloor \frac{C_{Fu}\tau_c}{K\tau_u} \right\rfloor\right). \tag{65}$$

where $\lfloor\cdot\rfloor$ is used to denote the floor function. Using this result, problem (58) can be rewritten as

$$\begin{aligned} &\max_{\boldsymbol{\nu}} \min_{k} \text{SINR}(\boldsymbol{\nu}, \tilde{\boldsymbol{\alpha}}_u^{\text{opt}}), \\ &\text{subject to } 0 \le \nu_k \le P_{uk} \quad \forall k \end{aligned} \tag{66}$$

or, in an equivalent form, as

$$\begin{aligned} &\max_{\boldsymbol{\nu},\, x} x \\ &\text{subject to } x \le \text{SINR}(\boldsymbol{\nu}, \tilde{\boldsymbol{\alpha}}_u^{\text{opt}}), \\ &\qquad 0 \le \nu_k \le P_{uk} \quad \forall k, \end{aligned} \tag{67}$$

which can be efficiently solved using a bisection search, where a convex feasibility problem is solved at each step [4].

#### 2) BCU-BASED CASE

Similar to the BAP-based case, the optimization problem for the BCU-based case can be formulated as

$$\begin{aligned} &\max_{(\boldsymbol{\nu},\, \tilde{\boldsymbol{\alpha}}_u,\, \tilde{\boldsymbol{\alpha}}_p)} \min_{k} \text{SINR}(\boldsymbol{\nu}, \tilde{\boldsymbol{\alpha}}), \\ &\text{subject to } 0 \le \nu_k \le P_{uk} \quad \forall k, \\ &\qquad B\frac{\tau_u}{\tau_c} b\left(\alpha_{um}\right) \le \epsilon_{C_{Fu}} C_{Fu} \quad \forall m, \\ &\qquad B\frac{\tau_p}{\tau_c} b\left(\alpha_{p_m}\right) \le (1-\epsilon_{C_{Fu}}) C_{Fu} \quad \forall m, \end{aligned} \tag{68}$$

where note that $\varrho_{mk}$ is a function of $\alpha_{p_m}$ and, thus, $\tilde{\boldsymbol{\alpha}} = [\tilde{\boldsymbol{\alpha}}_u^T \; \tilde{\boldsymbol{\alpha}}_p^T]^T$. The objective function is the achievable SINR of MS $k$ as calculated in Section III, where $A_{mk}(\tilde{\boldsymbol{\alpha}})$, $C_{kk'}(\tilde{\boldsymbol{\alpha}})$, and $D_k(\tilde{\boldsymbol{\alpha}})$ can be obtained using (60), (62) and (63), respectively, and

$$B_{kk'}(\tilde{\boldsymbol{\alpha}}) = \sum_{m=1}^{M} \tilde{\alpha}_{umk} \beta_{mk'} \varrho_{mk}. \tag{69}$$

The first constraint limits the power that can be allocated to MS $k$. The limited-capacity fronthaul in the UL introduces two constraints specifying that, out of the available UL fronthaul capacity $C_{Fu}$, only a part proportional to $\epsilon_{C_{Fu}}$ can be devoted to transmit the UL payload data (see (53)) and, correspondingly, only a part proportional to $(1-\epsilon_{C_{Fu}})$ can be allocated to the transmission of UL training-related information (see (54)). As the functions $b(\alpha_{um})$ and $b(\alpha_{p_m})$ are not continuous and can only take non negative integer values, the parameter $\epsilon_{C_{Fu}}$ can only be optimized by analyzing all the possible combinations of $b(\alpha_{um})$ and $b(\alpha_{p_m})$.

Again, the achievable rates monotonically increase with the quality of the quantizers used in both the UL training and payload transmission phases. Consequently, under optimal conditions, the number of bits used to quantize the corresponding samples should be selected as large as possible while fulfilling the fronthaul capacity constraints, that is,

$$\begin{aligned} b\left(\alpha_{um}^{\text{opt}}\right) &= \left\lfloor \frac{\epsilon_{C_{Fu}} C_{Fu}\tau_c}{\tau_u} \right\rfloor, \\ b\left(\alpha_{pm}^{\text{opt}}\right) &= \left\lfloor \frac{(1-\epsilon_{C_{Fu}}) C_{Fu}\tau_c}{\tau_p} \right\rfloor \end{aligned} \tag{70}$$

or, equivalently,

$$\begin{aligned} \tilde{\alpha}_{umk}^{\text{opt}} = \alpha_{um}^{\text{opt}} &= b^{-1}\left(\left\lfloor \frac{\epsilon_{C_{Fu}} C_{Fu}\tau_c}{\tau_u} \right\rfloor\right), \\ \alpha_{pm}^{\text{opt}} &= b^{-1}\left(\left\lfloor \frac{(1-\epsilon_{C_{Fu}}) C_{Fu}\tau_c}{\tau_p} \right\rfloor\right). \end{aligned} \tag{71}$$

These *optimal* parameters can be used to rewrite optimization problem (68) as in both (66) and (67) by only substituting $\tilde{\boldsymbol{\alpha}}_u^{\text{opt}}$ with $\tilde{\boldsymbol{\alpha}}^{\text{opt}}$. Hence, this optimization problem can also be solved by using a bisection search algorithm [4].

### B. DOWNLINK POWER CONTROL AND QUANTIZATION

Similar to the UL case, max-min DL power allocation and quantization problems aim at finding the vectors of power control coefficients $\boldsymbol{\eta}$ and payload quantization accuracy parameters $\tilde{\boldsymbol{\alpha}}_d$ that jointly maximize the minimum of the achievable DL rates (or, equivalently, the minimum of the achievable DL SINRs) of all MSs while satisfying the average transmit power and DL fronthaul capacity constraints at each AP [4], [12], [14], [23], [27].

#### 1) BAP-BASED CASE

For the BAP-based AP-CPU functional split, the optimization problem can be mathematically formulated as in (72), as shown at the bottom of the next page, where the objective function is the achievable SINR of MS $k$ as calculated in Section III (see, (44), (46), (48), and (51)), and the first

constraint indicates that the power consumption at the $m$th AP, as calculated in (33), must be less than or equal to the available transmit power $P_{dm}$. The second constraint implies that the average rate used to transfer the payload data signal on the $m$th fronthaul link, as calculated in (57), must not be higher than the corresponding allocated fronthaul bandwidth. Note that in posing this problem we have assumed, as in the BAP-based UL scenario, that the available DL fronthaul capacity $C_{Fd}$ is equally split among the $K$ active MSs.

As in the UL case, the DL achievable user rates monotonically increase with the quality of the quantization process during the payload transmission phase. Hence, under optimal conditions, the number of bits used to quantize the corresponding samples should be selected as large as possible while fulfilling the fronthaul capacity constraints, that is,

$$b\left(\alpha_{dmk}^{\text{opt}}\right) = \left\lfloor \frac{C_{Fd}\tau_c}{K\tau_d} \right\rfloor, \tag{73}$$

or, equivalently,

$$\alpha_{dmk}^{\text{opt}} = \alpha_{dm}^{\text{opt}} = b^{-1}\left(\left\lfloor \frac{C_{Fd}\tau_c}{K\tau_d} \right\rfloor\right). \tag{74}$$

Using these parameters, optimization problem (72) can be reformulated as

$$\begin{aligned}
&\max_{\{\boldsymbol{\varsigma},\boldsymbol{\lambda}\}} \min_k \frac{\left(\sum_{m=1}^{M} \alpha_{dm}^{\text{opt}} \varsigma_{mk} \varrho_{mk}^{1/2}\right)^2}{\sum_{m=1}^{M}\sum_{k'=1}^{K} \varsigma_{mk'}^2 \kappa_{mkk'} + \sum_{\substack{k'=1\\k'\neq k}}^{K} \lambda_{kk'}^2 + \sigma_d^2}, \\
&\text{subject to } \sqrt{\frac{\pi}{4}} \sum_{m=1}^{M} \alpha_{dm}^{\text{opt}} \varsigma_{mk'} \varrho_{mk}^{1/2} \left|\boldsymbol{\varphi}_{k'}^H \boldsymbol{\varphi}_k\right| \leq \lambda_{kk'} \quad \forall k' \neq k, \\
&\sum_{k'=1}^{K} \varsigma_{mk'}^2 \leq P_{dm}/\alpha_{dm}^{\text{opt}} \quad \forall m, \\
&\varsigma_{mk} \geq 0 \quad \forall mk,
\end{aligned} \tag{75}$$

where we have introduced the slack variables $\lambda_{kk'}$ and have used the definitions $\varsigma_{mk} = \eta_{mk}^{1/2}$ and

$$\kappa_{mkk'} = \alpha_{dm}^{\text{opt}}\left(\beta_{mk} - \frac{\pi}{4}\alpha_{dm}^{\text{opt}}\varrho_{mk}\left|\boldsymbol{\varphi}_{k'}^H\boldsymbol{\varphi}_k\right|^2\right).$$

Problem (75) is a quasi-concave optimization program that can be expressed in an equivalent form as

$$\begin{aligned}
&\max_{\{\boldsymbol{\varsigma},\boldsymbol{\lambda},x\}} x \\
&\text{s. t. } \sqrt{x}\left\|\left[\boldsymbol{\mu}_{1k}\ldots\boldsymbol{\mu}_{Mk}\,\bar{\boldsymbol{\lambda}}_k\,\sigma_d\right]\right\| \leq \sum_{m=1}^{M} \varsigma_{mk}\alpha_{dm}^{\text{opt}}\varrho_{mk}^{1/2} \quad \forall k, \\
&\sqrt{\frac{\pi}{4}}\sum_{m=1}^{M}\varsigma_{mk'}\alpha_{dm}^{\text{opt}}\varrho_{mk}^{1/2}\left|\boldsymbol{\varphi}_{k'}^H\boldsymbol{\varphi}_k\right| \leq \lambda_{kk'} \quad \forall k' \neq k \\
&\left\|\boldsymbol{\varsigma}_m\right\| \leq \sqrt{P_{dm}/\alpha_{dm}^{\text{opt}}} \quad \forall m, \\
&\varsigma_{mk} \geq 0 \quad \forall m,\ k,
\end{aligned} \tag{76}$$

with $\boldsymbol{\varsigma}_m = [\varsigma_{m1}\ldots\varsigma_{mK}]$, $\boldsymbol{\mu}_{mk} = \left[\varsigma_{m1}\kappa_{mk1}^{1/2}\ldots\varsigma_{mK}\kappa_{mkK}^{1/2}\right]$ and $\bar{\boldsymbol{\lambda}}_k = \left[\lambda_{k1}\ldots\lambda_{k(k-1)}\lambda_{k(k+1)}\ldots\lambda_{kK}\right]$. Problem (76) is a second order cone (SOC) program that can be efficiently solved by using a conventional iterative bisection search algorithm. Specific details on the optimality, complexity and feasibility of these algorithms were fully commented by Ngo *et al.* in the seminal paper [4].

### 2) BCU-BASED CASE

Analogously to the BAP-based case, the optimization problem for the BCU-based AP-CPU functional split can be expressed as in (77), as shown at the bottom of the next page. Moreover, once again, as the DL achievable user rates monotonically increase with the quality of the quantization process, the number of bits used to quantize the corresponding samples, under optimal conditions, must be as large as possible while fulfilling the fronthaul capacity constraints and, consequently,

$$b\left(\alpha_{dm}^{\text{opt}}\right) = \left\lfloor \frac{C_{Fd}\tau_c}{\tau_d} \right\rfloor, \tag{78}$$

$$\begin{aligned}
&\max_{(\boldsymbol{\eta},\,\boldsymbol{\alpha}_d)} \min_k \frac{\frac{\pi}{4}\left(\sum_{m=1}^{M}\alpha_{dmk}\eta_{mk}^{1/2}\varrho_{mk}^{1/2}\right)^2}{\sum_{k'=1}^{K}\sum_{m=1}^{M}\alpha_{dmk'}\eta_{mk'}\left(\beta_{mk} - \frac{\pi}{4}\alpha_{dmk'}\varrho_{mk}\left|\boldsymbol{\varphi}_{k'}^H\boldsymbol{\varphi}_k\right|^2\right) + \frac{\pi}{4}\sum_{k'\neq k}\left(\sum_{m=1}^{M}\alpha_{dmk'}\eta_{mk'}^{1/2}\varrho_{mk}^{1/2}\right)^2\left|\boldsymbol{\varphi}_{k'}^H\boldsymbol{\varphi}_k\right|^2 + \sigma_d^2}, \\
&\text{subject to } \sum_{k=1}^{K}\alpha_{dmk}\eta_{mk} \leq P_{dm} \quad \forall m, \\
&B\frac{\tau_d}{\tau_c}b\left(\alpha_{dmk}\right) \leq \frac{C_{Fd}}{K} \quad \forall mk, \\
&\eta_{mk} \geq 0 \quad \forall mk
\end{aligned} \tag{72}$$

or, equivalently,

$$\alpha_{dm}^{\text{opt}} = b^{-1}\left(\left\lfloor \frac{C_{Fd}\tau_c}{\tau_d} \right\rfloor\right). \tag{79}$$

Using these *optimal* quantization quality indicators, optimization problem (77) can be straightforwardly rewritten as in both (75) and (76), allowing it to be solved by means of a bisection search algorithm.

## VI. NUMERICAL RESULTS AND DISCUSSIONS

In this section, numerical results are provided to quantitatively assess the performance of fronthaul-constrained cell-free massive MIMO in terms of the max-min per-user achievable rate. A special emphasis is put on the performance comparison between the proposed CPU-AP functional splits at the physical layer (i.e., BAP and BCU strategies) under different fronthaul operational conditions. The cell-free scenario under evaluation replicates the one typically used in the most relevant literature on this topic (see, for instance, [4], [12], [14], [23], [27], [28]). In particular, the $M$ APs and $K$ MSs are uniformly distributed at random within a square coverage area with a side equal to $D$. Boundary effects are avoided by wrapping around this square area at the edges. Nevertheless, as suggested by Björnson and Sanguinetti in [33], use is made of the large-scale propagation loss model

$$\zeta_{mk}[dB] = -30.5 - 36.7\log_{10} d_{mk}, \tag{80}$$

where $d_{mk}$ is the distance between the $m$th AP and MS $k$ (measured in meters and computed by taking into account the wrap-around implementation and the heights of both the AP and MS). The shadowing component $\chi_{mk}$ is modelled as a correlated log-normal random variable with variance $\sigma_\chi^2$ whose spatial correlation model is described in [4, (54)-(55)], with the decorrelation distance set to $d_{\text{decorr}} = 9$ m and the parameter $\delta$ set to 0.5. The default parameters used to set-up the simulation scenarios under evaluation in the following subsections will be those summarized in Table 2.

**TABLE 2. Summary of default simulation parameters.**

| Parameter | Value |
|---|---|
| Carrier frequency: $f_0$ | 2 GHz |
| Bandwidth: $B$ | 20 MHz |
| Side of the square coverage area: $D$ | 1000 m |
| Number of APs: $M$ | 100 APs |
| AP antenna height: $h_{AP}$ | 15 m |
| MS antenna height: $h_{MS}$ | 1.65 m |
| Noise figure (AP and MS): $NF$ | 9 dB |
| Available power at the AP: $P_d$ | 200 mW |
| Available power at the MS: $P_u = P_p$ | 100 mW |
| Coherence interval length: $\tau_c$ | 200 samples |
| Training phase length: $\tau_p$ | 20 samples |
| Payload phase length: $\tau_d = \tau_u$ | $(\tau_c - \tau_p)/2$ samples |

### A. IMPACT OF THE PILOT ALLOCATION PROCESS

Our aim in this subsection is to benchmark the performance of different pilot allocation strategies against an *ideal* scheme. The *ideal* pilot allocation strategy assumes that, irrespective of the number of active MSs in the network, it is always possible to allocate an orthogonal pilot to each of them. In other words, we are basically disregarding the pilot contamination effects and hence obtaining an upper bound on the potential performance any pilot allocation policy can offer. The practical allocation schemes that we benchmark against the ideal one are[4]:

- *Random pilot allocation (RPA)*: In this case, irrespective of the number of active MSs in the network, each MS is allocated a pilot sequence randomly selected from the set of $\tau_p$ orthogonal pilot sequences.
- *Naive*: This strategy is based on a slight modification of the pure RPA scheme. For those cases in which $K \leq \tau_p$, each MS is allocated an orthogonal pilot sequence. For those cases in which $K > \tau_p$, instead, there are $\tau_p$ MSs that are allocated the $\tau_p$ orthogonal pilot sequences, and each of the remaining $K - \tau_p$ MSs is allocated a pilot sequence randomly selected from the pool of available orthogonal ones.

[4]Note that implementing an optimal pilot allocation strategy is an NP-hard problem. Even the conceptually simple greedy algorithm proposed by Ngo *et al.* in [4], which can only be implemented by giving priority to the performance of either the UL or the DL, has a very high implementation complexity that makes it difficult to be implemented in practice when the number of active MSs in the network is relatively large.

$$\max_{(\boldsymbol{\eta},\,\boldsymbol{\alpha}_d)} \min_k \frac{\frac{\pi}{4}\left(\sum_{m=1}^{M}\alpha_{dm}\eta_{mk}^{1/2}\varrho_{mk}^{1/2}\right)^2}{\sum_{k'=1}^{K}\sum_{m=1}^{M}\alpha_{dm}\eta_{mk'}\left(\beta_{mk} - \frac{\pi}{4}\alpha_{dm}\varrho_{mk}\left|\boldsymbol{\varphi}_{k'}^H\boldsymbol{\varphi}_k\right|^2\right) + \frac{\pi}{4}\sum_{k'\neq k}\left(\sum_{m=1}^{M}\alpha_{dm}\eta_{mk'}^{1/2}\varrho_{mk}^{1/2}\right)^2\left|\boldsymbol{\varphi}_{k'}^H\boldsymbol{\varphi}_k\right|^2 + \sigma_d^2},$$

$$\text{subject to } \alpha_{dm}\sum_{k=1}^{K}\eta_{mk} \leq P_{dm} \quad \forall\, m,$$

$$B\frac{\tau_d}{\tau_c}b\left(\alpha_{dm}\right) \leq C_{Fd} \quad \forall\, m,$$

$$\eta_{mk} \geq 0 \quad \forall\, mk \tag{77}$$

- *Balanced random pilot allocation (BRPA)*: In order to avoid unnecessary overuse of any of the available orthogonal pilot sequences, in this scheme each MS is allocated a pilot sequence that is sequentially and cyclically selected from the ordered set of $\tau_p$ orthogonal pilots.
- *Dissimilarity cluster-based pilot assignment (DCPA)*: Given any active MS $k$, the CPU has a perfect knowledge of vector $\boldsymbol{\beta}_k = [\beta_{1k} \ldots \beta_{Mk}]^T$ containing the large-scale propagation gains of the channels linking this MS to the $M$ APs. Vector $\boldsymbol{\beta}_k$ can be regarded as an effective *fingerprint* characterizing the location of MS $k$. The DCPA strategy, first proposed in [27], uses the so-called *cosine similarity* measure to ensure that pilot sequences are only reused, in a balanced manner, by MSs showing the most *dissimilar* large-scale propagation patterns to the APs.

The average max-min rate per-user *versus* the number of active MSs is presented in Fig. 2 for each of these pilot allocation strategies and for both the DL and the UL. All results have been obtained assuming the use of the BCU-based CPU-AP functional split, the default system parameters described in Table 2, a DL fronthaul link with a capacity of $C_{Fd} = 33$ Mbps, and an UL fronthaul link with $C_{Fu} = 33$ Mbps and $\epsilon_{C_{Fu}} = 9/11$. Using (70) and (78) it can be easily shown that these parameters correspond to a case in which all the quantizers implemented during the training and payload transmission phases use exactly 3 bit/sample. A first result that is worth emphasizing is that, although the achievable max-min user rates in the DL are much higher than those achievable in the UL,[5] the behavioral patterns of this metric are virtually identical in both links. As expected, the average max-min user rates obtained assuming the use of an *ideal* pilot allocation scheme constitute an upper bound on the performance provided by any other pilot allocation schemes. The pure RPA scheme is clearly outperformed by the *naive*, the BRPA and the DCPA strategies irrespective of the number of active MSs in the network. In fact, even for those cases in which $K \leq \tau_p$ (in this setup $\tau_p = 20$ samples), the RPA scheme cannot guarantee the absence of pilot reuse. Furthermore, for those cases in which $K > \tau_p$, this pilot allocation strategy does not prevent situations in which a given pilot is allocated to a large number of MSs and/or to MSs exhibiting very similar large-scale propagation patterns to the APs. Thus, irrespective of the number of active MSs in the network, RPA schemes present a high probability of having one or more users suffering from high levels of pilot contamination, with the consequent reduction of the achievable max-min user rate. All the other strategies completely eliminate this effect for those cases in which $K \leq \tau_p$ thanks to the use of orthogonal pilot allocation. For those cases in which $K > \tau_p$, however, pilot contamination effects are unavoidable and the *naive*, BRPA and DCPA schemes rely on different procedures to diminish its effects. In particular, under the *naive* scheme there are $K - \tau_p$ MSs that are allocated a pilot sequence randomly selected from the pool of available orthogonal ones and, hence, although it shows a better performance than the pure RPA scheme, as the value of $K$ increases both strategies tend to exhibit the same limitations in terms of pilot contamination avoidance. Turning our attention to the BRPA and DCPA strategies, it can be observed that, on the one hand, the BRPA scheme clearly outperforms the *naive* one by only precluding the unbalanced reuse of pilots. On the other hand, at the cost of a negligible increase in complexity, the DCPA strategy not only provides a balanced pilot reuse, but it also guarantees that MSs sharing a given pilot exhibit very dissimilar large-scale propagation patterns, thus providing a slight, yet perceptible, performance advantage over the BRPA scheme. Results presented from this point onwards will be obtained assuming the use of the DCPA strategy.

[5]Note that DL optimization involves the setting of $MK$ power allocation coefficients whereas the UL counterpart only requires the optimization of $K$ parameters. The larger number of degrees-of-freedom in the DL is the main cause of its superior performance. Although more sophisticated precoding/decoding techniques could be applied in the UL, they would certainly not alter the general conclusions drawn from results presented in this paper.

### B. COMPARING CPU-AP FUNCTIONAL SPLITS UNDER THE EFFECTS OF FRONTHAUL CAPACITY CONSTRAINTS

Given the capacity constraints of the fronthaul links, it makes one wonder whether it is most convenient to implement precoding processes at the APs or at the CPU. Looking at the description of BCU- and BAP-based functional splits in Fig. 1, and as the CSI varies at the pace of $\tau_p$ and data varies at a much higher pace, our intuition says that the fronthaul capacity constraints should be more detrimental for the BAP-based CPU-AP functional split than for the BCU-based one. The average max-min rate per-user results presented in Figs. 3 and 4 clearly confirm our intuition in scenarios with different number of active MSs and different fronthaul capacity constraints. In particular, it can be observed comparing these figures that both CPU-AP functional splits provide

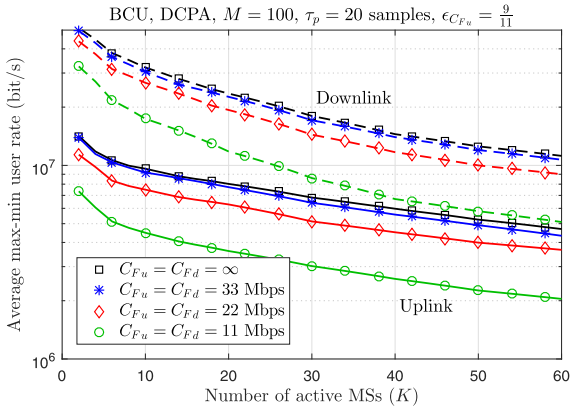
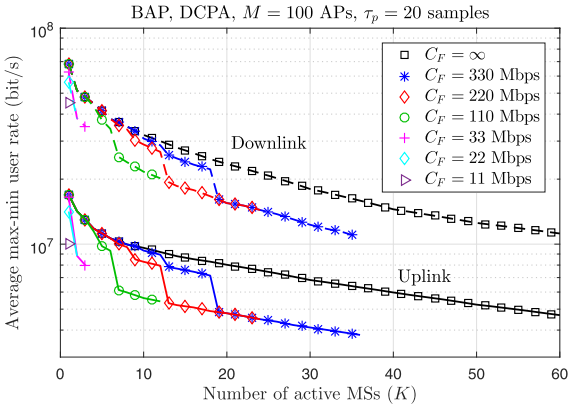
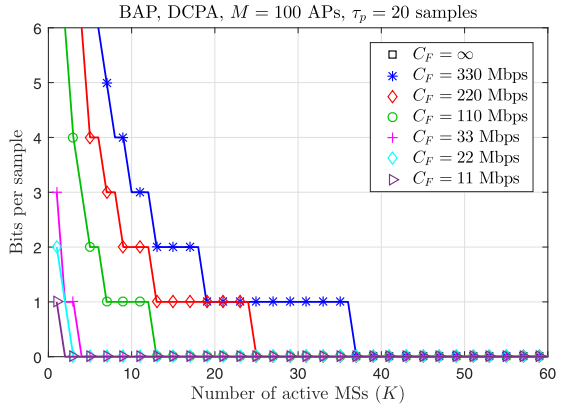



**FIGURE 3.** Average max-min rate per-user versus the number of active MSs, assuming the use of the BCU-based CPU-AP functional split and different fronthaul capacity constraints (DCPA, BCU, $M = 100$ APs, $\tau_p = 20$ samples, $\epsilon_{C_{Fu}} = 9/11$).




(a) Average max-min user rate




(b) Optimal number of bits per sample

**FIGURE 4. Average max-min rate per-user and optimal number of bits per sample versus the number of active MSs, assuming the use of the BAP-based CPU-AP functional split and different fronthaul capacity constraints (DCPA, BAP, $M = 100$ APs, $\tau_p = 20$ samples).**

exactly the same performance under unconstrained capacity fronthaul links, which acts as an upper-bound on the performance that can be obtained in any capacity-constrained fronthaul scenario. As the fronthaul capacity decreases, however, the max-min per-user rate experienced under BCU- and BAP-based strategies suffer from very dissimilar quantization distortion-based effects.

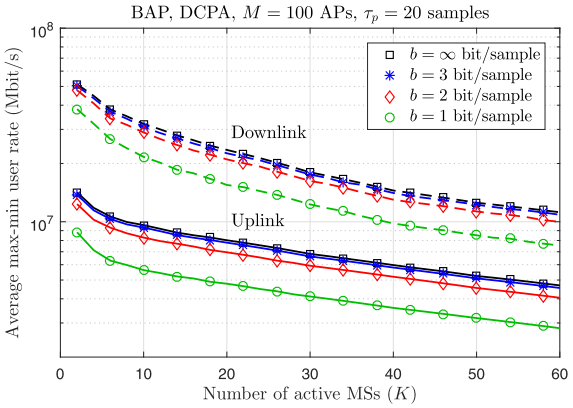
The fronthaul capacity consumption of a BCU-based system does not depend on the number of active MSs in the system (see (70) and (78)) and, consequently, as it can be observed in Fig. 3, the performance provided by this CPU-AP functional split scales adequately with $K$. For instance, even though constraining the fronthaul capacity to $C_{Fu} = C_{Fd} = C_F = 33$, 22 or 11 Mbps, with $\tau_c = 200$ samples, $\tau_p = 20$ samples, and $\epsilon_{C_{Fu}} = 9/11$, requires the use of $b = 3$, 2, or 1 bit/sample low-resolution ADCs, respectively, the cell-free massive MIMO network can still provide a fairly good service to a very large number of active MSs. Obviously, the price to pay is a non-negligible decrease in the max-min per-user rate. Interestingly, using a fronthaul capacity $C_{Fu} = C_{Fd} = C_F \geq 44$ Mbps allows for the use of ADCs with $b \geq 4$ bits/sample that provide a virtually negligible performance degradation with respect to the infinite fronthaul capacity scenario.

The fronthaul capacity consumption of the BAP-based CPU-AP functional split, on the contrary, has a proportional dependency on the number of active MSs in the system (see (64) and (73)) and, inevitably, as it can be observed in Fig. 4a, the performance provided by this CPU-AP functional split suffers dramatic consequences as the number of active MSs increases. As shown in Fig. 4b, constraining the fronthaul capacity to $C_{Fu} = C_{Fd} = C_F = 33$, 22 or 11 Mbps, with $\tau_c = 200$ samples and $\tau_p = 20$ samples, allows providing service up to only $K = 3$, 2 or 1 MSs, respectively, assuming the use of a 1-bit ADC. Even when considering a ten-fold increase in fronthaul capacity, that is, constraining the fronthaul capacity to $C_{Fu} = C_{Fd} = C_F = 330$, 220 or 110 Mbps, the BAP-based scheme is nowhere near the performance achieved by the BCU-based CPU-AP functional split. In fact, under this high-bandwidth scenario, the BCU-based strategy could easily rely on the use of high-resolution ADCs providing a negligible performance degradation with respect to the infinite-capacity fronthaul links. The BAP-based scheme, however, would require the use of variable-resolution ADCs, as shown in Fig. 4b, with the consequent performance loss when using low-resolution quantization processes and without guarantying the possibility to provide service to all MSs in highly populated scenarios.

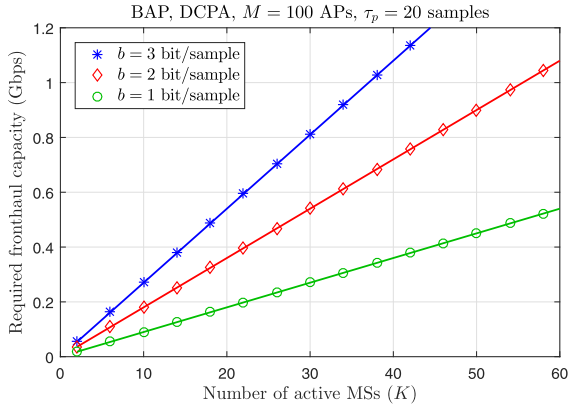
Comparing the performance results presented in Fig. 3 and those presented in Fig. 5a is just another way of looking at the same problem. Results presented in these figures correspond to scenarios in which the low-resolution ADCs are considered using $b = 3$, 2 or 1 bits per sample. Remarkably, as the BAP-based CPU-AP functional split does not need to compress the CSI, the performance provided by this strategy, assuming the use of the same quantization resolution, is better than that provided by the CPU-based scheme thanks to the availability of undistorted MMSE channel estimates. The key point to realize, however, is that the BCU-based strategy is able to obtain the max-min per-user rates shown in Fig. 3 by consuming very low fronthaul capacities that, furthermore, do not depend on the number of active MSs in the network. The BAP-based scheme, in contrast, consumes a large amount of fronthaul capacity that, as shown in Fig. 5b, increases linearly with the number of MSs served by the cell-free massive MIMO network.

### C. OPTIMIZING THE PARAMETER $\epsilon_{C_{Fu}}$

Results presented in this subsection will be obtained assuming the use of the BCU-based CPU-AP functional split. Under this functional split, a fraction of the UL fronthaul capacity $C_{Fu}$, proportional to $\epsilon_{C_{Fu}}$, is allocated to the UL payload data transmission phase, and a fraction proportional to $(1 - \epsilon_{C_{Fu}})$ is allocated to the UL training phase. Results presented in Fig. 6, which have been obtained using $M = 100$ APs, $\tau_p = 20$ samples and $C_{Fu} = C_{Fd} = 33$ Mbps, serve to understand the impact produced by modifying the parameter $\epsilon_{C_{Fu}}$ on the UL and DL achievable max-min per-user

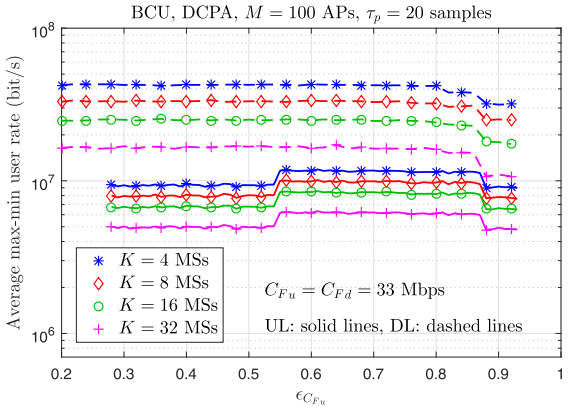



(a) Average max-min user rate

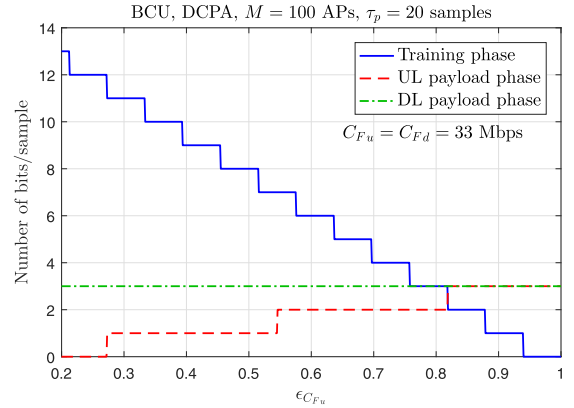



(b) Required fronthaul capacity

**FIGURE 5.** Average max-min rate per-user and required fronthaul capacity versus the number of active MSs, assuming the use of the BAP-based CPU-AP functional split and different ADC resolutions (DCPA, BAP, $M = 100$ APs, $\tau_p = 20$ samples).



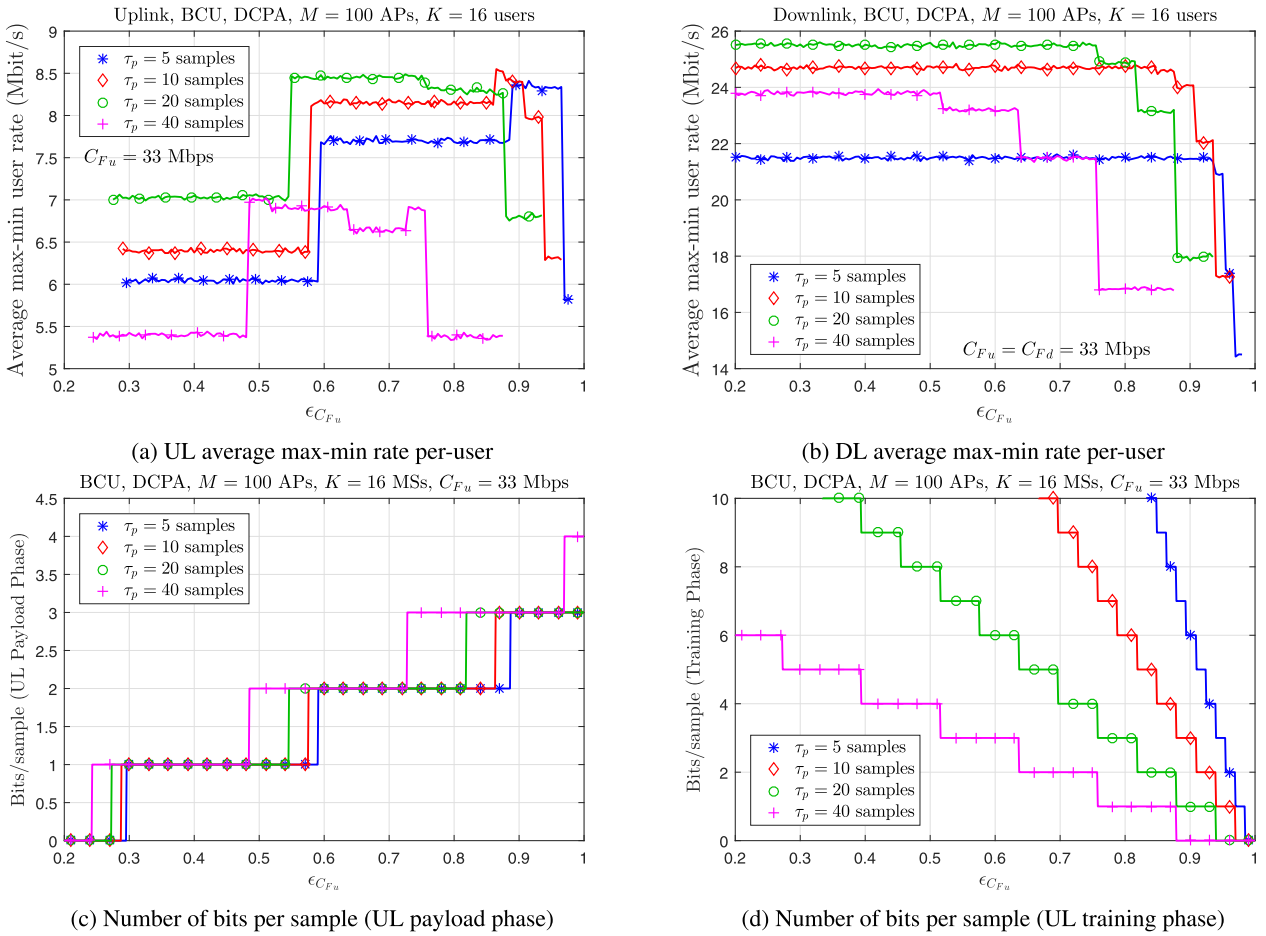

(a) Average max-min rate per-user versus $\epsilon_{C_{Fu}}$




(b) Number of bits per sample versus $\epsilon_{C_F}$

**FIGURE 6.** Average max-min rate per-user and number of bits per sample versus $\epsilon_{C_{Fu}}$ assuming the use of the BCU functional split and with the number of active MSs as parameter (DCPA, $M = 100$ APs, $\tau_p = 20$ samples, $C_{Fu} = C_{Fd} = 33$ Mbps).

rates in cell-free massive MIMO networks serving different amounts of MSs. Note that in order to better understand the behavior of the average max-min per-user rate as a function of $\epsilon_{C_{Fu}}$, presented in Fig. 6a, the number of bits per sample used in the quantization processes implemented during the different transmission phases (UL training phase, UL payload phase and DL payload phase) are also plotted in Fig. 6b. On the one hand, for very high values of the parameter $\epsilon_{C_{Fu}}$, the amount of UL fronthaul capacity allocated to convey the CSI generated during the UL training phase is very small and, hence, the optimal number of quantization bits per sample must also necessarily be small, with the consequent increase in quantization noise power resulting in a clear decrease of the achievable max-min user rates in both UL/DL payload transmission phases. In fact, with $\tau_c = 200$ samples and $\tau_p = 20$ samples, the ADC used during the UL training phase cannot support values of $\epsilon_{C_{Fu}} \geq 0.94$, and can only provide one bit per sample for $0.879 \leq \epsilon_{C_{Fu}} < 0.94$. On the other hand, for very low values of the parameter $\epsilon_{C_{Fu}}$, the APs are allocated a large amount of fronthaul capacity during the UL training phase allowing to convey a virtually quantization noise-free CSI to the CPU. The price to pay for this *ideal* quantization noise-free CSI is a decrease in the amount of UL fronthaul capacity allocated to the UL payload data transmission phase that, for obvious reasons, results in a decrease in the optimal achievable max-min rates per-user. In fact, with the system parameters assumed in this scenario, the ADC used during the UL payload transmission phase cannot support values of $\epsilon_{C_{Fu}} \leq 0.272$, and can only provide one bit per sample for $0.273 \leq \epsilon_{C_{Fu}} < 0.546$. As a consequence of this fairly predictable behavior, there is a particular range of values of the parameter $\epsilon_{C_{Fu}}$ for which the UL achievable max-min per-user rate is optimal that, for this particular scenario is $0.546 \leq \epsilon_{C_{Fu}} < 0.576$ (note that in this range the optimal number of bits per sample used during the UL training and payload transmission phases is equal to 6 and 2, respectively). It is worth stressing that this optimal range of values has been obtained by only observing the behavior of the UL performance results. This is because the number of fronthaul samples (channel uses) allocated to the DL payload transmission phase is independent of $\epsilon_{C_{Fu}}$ and so it is the optimal number of bits per sample used by the corresponding low-resolution ADC (3 bits/sample for the scenario under evaluation). Hence, the variations of the




(a) UL average max-min rate per-user

(b) DL average max-min rate per-user

(c) Number of bits per sample (UL payload phase)

(d) Number of bits per sample (UL training phase)

**FIGURE 7.** Average max-min rate per-user and number of bits per sample versus $\epsilon_{C_{Fu}}$ assuming the use of the BCU functional split and with the UL training phase duration $\tau_p$ as parameter (DCPA, $M = 100$ APs, $K = 16$ MSs, $C_{Fu} = C_{Fd} = 33$ Mbps).

DL average max-min per-user rate as a function of $\epsilon_{C_{Fu}}$ only depend on the quality of the quantized CSI provided during the UL training phase. Thus, notwithstanding that even though the DL achievable max-min per-user rates are virtually constant for any $\epsilon_{C_{Fu}} \leq 0.697$ (using an ADC with more than 4 bit/sample is practically equivalent to the use of an ideal ADC), they decrease in a staggered way as the quality of the UL training phase quantizer decreases with the increase of $\epsilon_{C_{Fu}}$.

In order to gain a rather complete picture of the impact the modification of $\epsilon_{C_{Fu}}$ may have on the system performance, Fig. 7 shows the average max-min rate per-user and number of bits per sample versus $\epsilon_{C_{Fu}}$ assuming a cell-free massive MIMO network serving $K = 16$ MSs and using different values of the UL training phase duration $\tau_p$. As in Fig. 6, we can observe that for the UL case there is always a range of values of the parameter $\epsilon_{C_{Fu}}$ for which the achievable max-min per-user rate is optimal. Remarkably, the specific optimal range of $\epsilon_{C_{Fu}}$ depends on the particular value of $\tau_p$. This is basically due to the dependence of the optimal number of quantization bits per sample on $\tau_p$ and $\epsilon_{C_{Fu}}$. In particular, the higher the value of $\tau_p$, the larger the amount of CSI to be conveyed from the APs to the CPU and thus, the lower the optimal range of values of $\epsilon_{C_{Fu}}$. Specifically, as it can be observed in Fig. 7a, the optimal values for the parameter $\epsilon_{C_{Fu}}$ are $0.887 \leq \epsilon_{C_{Fu}} < 0.893$ (where the UL training phase quantizer is using 6 bits/sample and the UL payload transmission phase quantizer is using 3 bits/sample), for $\tau_p = 5$ samples, and $0.485 \leq \epsilon_{C_{Fu}} < 0.515$ (where the UL training phase quantizer is using 4 bits/sample and the UL payload transmission phase quantizer is using 2 bits/sample), for $\tau_p = 40$ samples. Again, as shown in Fig. 7b, the DL achievable max-min per-user rates are virtually constant for low values of $\epsilon_{C_{Fu}}$ (i.e, in the range where the UL training phase ADC can use more than 4 bit/sample), but they decrease in a staggered way as the quality of the quantized CSI decreases due to the increase of $\epsilon_{C_{Fu}}$.

## VII. CONCLUSION

The max-min per-user rate performance provided by fronthaul-constrained cell-free massive MIMO networks using low-resolution ADCs has been analyzed under different CPU-AP functional splits. The proposed analytical framework considers the use of matched filtering in the

UL and normalized conjugate beamforming in the DL. The impact of using low-resolution ADCs to transmit over capacity-constrained fronthaul links has been addressed by resorting to the use of a linear additive quantization noise model that is based on the Bussgang decomposition and is sometimes referred to as the AQNM. The derivation of mathematically tractable closed-form expressions for both the per-user achievable rates and the fronthaul bandwidth consumption has allowed posing max-min power allocation and fronthaul quantization optimization problems that have been solved using standard convex optimization tools. Results have shown that, under capacity-constrained fronthaul links, the *classical* cell-free massive MIMO networks with single-antenna APs and using BAP-based CPU-AP functional splits are clearly outperformed, in terms of max-min user rate performance, by the BCU-based CPU-AP functional splits. In contrast, if the limiting factor is the resolution of the ADCs used to quantize the samples to be transmitted on the fronthaul links, the preferred CPU-AP functional splits are those in which the baseband processing is performed at the APs. It has also been shown that the UL fronthaul capacity fraction allocated to share the CSI among APs and CPU should be adapted as a function of the UL training phase length. Furthermore, numerical results indicate that the suboptimal DCPA pilot allocation scheme, which is based on the idea of clustering by dissimilarity, outperforms the pure random, *naive* and balanced random algorithms and approaches the performance provided by an *ideal* strategy at a fraction of the complexity burden associated with the NP-hard optimal schemes. Future work on this topic should be devoted to analyze the impact the fronthaul capacity constraints may have on the energy efficiency of cell-free massive MIMO networks using low-resolution ADCs by taking into account the power consumption of all the signal processing units in the system, and the possible use of multiple-antenna APs. It would also be interesting to explore the design of optimal resource allocation strategies between UL and DL, and the effects the use of adaptive training phases and a non-uniform distribution of MSs and/or APs might have on the performance of these networks. The impact fronthaul constraints may have in the context of cell-free massive MIMO architectures relying on centralized precoding strategies (i.e., ZF, MMSE) constitutes one more interesting research thread. Finally, further work should be devoted to analyze the impact the use of more accurate quantization noise models may have on the performance of cell free massive MIMO netwoks using low resolution ADCs.

## APPENDIX A
## CALCULATION OF THE STATISTICAL TERMS IN THE ACHIEVABLE RATES

### A. COMPUTATION OF $\mathbb{E}\{g_{mk}\varpi_{lmk}\}$

The propagation channel between AP $m$ and MS $k$ can be expressed as $g_{mk} = \hat{g}_{mk} + \tilde{g}_{mk}$, where $\hat{g}_{mk}$ is the channel estimation and $\tilde{g}_{mk}$ is the estimation error. Owing to the properties of MMSE estimation, $\hat{g}_{mk}$ and $\tilde{g}_{mk}$ are independent and thus, using either the MF detector in the UL or the NCB precoder in the DL, the expectation $\mathbb{E}\{g_{mk}\varpi_{lmk}\}$ for these particular cases can be written as

$$\begin{aligned}\mathbb{E}\{g_{mk}\varpi_{umk}\} &= \mathbb{E}\left\{\left(\hat{g}_{mk}+\tilde{g}_{mk}\right)\hat{g}^*_{mk}\right\}\\ &= \mathbb{E}\left\{\left|\hat{g}_{mk}\right|^2\right\} = \varrho_{mk}.\end{aligned} \tag{81}$$

or

$$\begin{aligned}\mathbb{E}\{g_{mk}\varpi_{dmk}\} &= \mathbb{E}\left\{\left(\hat{g}_{mk}+\tilde{g}_{mk}\right)\frac{\hat{g}^*_{mk}}{\left|\hat{g}_{mk}\right|}\right\}\\ &= \mathbb{E}\left\{\left|\hat{g}_{mk}\right|\right\} = \frac{\sqrt{\pi}}{2}\sqrt{\varrho_{mk}}.\end{aligned} \tag{82}$$

respectively, where it has been taken into account that $\hat{g}_{mk} \sim \mathcal{CN}(0, \varrho_{mk})$.

### B. COMPUTATION OF $\mathrm{Var}\{g_{mk}\varpi_{lmk}\}$

Assuming the use of a MF detector the variance of $g_{mk}\varpi_{umk}$ can be obtained as

$$\begin{aligned}\mathrm{Var}\{g_{mk}\varpi_{umk}\} = \mathrm{Var}\left\{g_{mk}\hat{g}^*_{mk}\right\} &= \mathbb{E}\left\{\left|\hat{g}_{mk}\right|^4\right\}\\ &+ \mathbb{E}\left\{\left|\hat{g}_{mk}\right|^2\right\}\mathbb{E}\left\{\left|\tilde{g}_{mk}\right|^2\right\} - \left(\mathbb{E}\left\{\left|\hat{g}_{mk}\right|^2\right\}\right)^2\\ &= 2\varrho^2_{mk} + \varrho_{mk}(\beta_{mk}-\varrho_{mk}) - \varrho^2_{mk} = \varrho_{mk}\beta_{mk}.\end{aligned} \tag{83}$$

Under the NCB precoding rule, the variance of $g_{mk}\varpi_{dmk}$ can be straightforwardly obtained as

$$\begin{aligned}\mathrm{Var}\{g_{mk}\varpi_{dmk}\} &= \mathrm{Var}\left\{g_{mk}\frac{\hat{g}^*_{mk}}{\left|\hat{g}_{mk}\right|}\right\}\\ &= \mathbb{E}\left\{\left|g_{mk}\right|^2\right\} - \left(\mathbb{E}\left\{\left|\hat{g}_{mk}\right|\right\}\right)^2 = \beta_{mk} - \frac{\pi}{4}\varrho_{mk}.\end{aligned} \tag{84}$$

### C. COMPUTATION OF $\mathbb{E}\left\{\left|\mathrm{UI}_{ukk'}\right|^2\right\}$

When calculating the terms $\mathbb{E}\left\{g_{mk'}g^*_{nk'}\varpi_{umk}\varpi^*_{unk}\right\}$ and $\mathbb{E}\left\{g_{mk}g^*_{nk}\varpi_{dmk'}\varpi^*_{dnk'}\right\}$, we have to distinguish between two particular cases. First, the case in which $m = n$, corresponding to an inter-user interference term received at the same AP as the useful term in the UL or to an inter-user interference term originating from the same AP as the useful term in the DL. Second, the case in which $m \neq n$, corresponding to an inter-user interference term received at a different AP than the useful term in the UL or to an inter-user interference term originating from a different AP than the useful term in the DL. Furthermore, two different situations must be examined in each of these cases. First, the situation in which both the tagged MS $k$ and the interfering MS $k'$ do not use the same pilot code (i.e., $\boldsymbol{\varphi}_k \neq \boldsymbol{\varphi}_{k'}$) and, consequently, $g_{mk}$ and $\hat{g}_{mk'}$ are uncorrelated. Second, the case in which both the tagged MS $k$ and the interfering MS $k'$ have been allocated the same pilot code (i.e., $\boldsymbol{\varphi}_k = \boldsymbol{\varphi}_{k'}$) and, hence, $\xi_{mk'} = \xi_{mk}$ and $\hat{g}_{mk'} = (\beta_{mk'}/\beta_{mk})\hat{g}_{mk}$ (i.e., except for a real multiplicate

constant, the corresponding channel estimations at the $m$th AP are exactly the same).

*Case* $m = n$: The inter-user interference expectation for $m = n$ can be obtained as

$$\begin{aligned}
&\mathbb{E}\left\{g_{mk'}g_{nk'}^{*}\varpi_{umk}\varpi_{unk}^{*}\right\}\\
&\quad=\mathbb{E}\left\{\left|g_{mk'}\hat{g}_{mk}^{*}\right|^{2}\right\}\\
&\quad=\begin{cases}\left(\beta_{mk}^{2}/\beta_{mk'}^{2}\right)\mathbb{E}\left\{\left|g_{mk'}\hat{g}_{mk'}^{*}\right|^{2}\right\}, & \boldsymbol{\varphi}_{k}=\boldsymbol{\varphi}_{k'}\\ \mathbb{E}\left\{\left|g_{mk'}\right|^{2}\right\}\mathbb{E}\left\{\left|\hat{g}_{mk}\right|^{2}\right\}, & \boldsymbol{\varphi}_{k}\neq\boldsymbol{\varphi}_{k'}\end{cases}\\
&\quad=\begin{cases}\left(\beta_{mk}^{2}/\beta_{mk'}^{2}\right)\varrho_{mk'}\left(\varrho_{mk'}+\beta_{mk'}\right), & \boldsymbol{\varphi}_{k}=\boldsymbol{\varphi}_{k'}\\ \varrho_{mk}\beta_{mk'}, & \boldsymbol{\varphi}_{k}\neq\boldsymbol{\varphi}_{k'}\end{cases}\\
&\quad=\varrho_{mk}\beta_{mk'}+\frac{\beta_{mk}^{2}}{\beta_{mk'}^{2}}\varrho_{mk'}^{2}\left|\boldsymbol{\varphi}_{k'}^{H}\boldsymbol{\varphi}_{k}\right|^{2},\quad (m\neq n)
\end{aligned}\tag{85}$$

for the UL, and simplifies to

$$\begin{aligned}
\mathbb{E}\left\{g_{mk}g_{nk}^{*}\varpi_{dmk'}\varpi_{dnk'}^{*}\right\}&=\mathbb{E}\left\{\left|g_{mk}\frac{\hat{g}_{mk'}^{*}}{|\hat{g}_{mk'}|}\right|^{2}\right\}\\
&=\mathbb{E}\left\{|g_{mk}|^{2}\right\}=\beta_{mk},\quad (m=n),
\end{aligned}\tag{86}$$

for the DL, irrespective of the pilot codes used by both the tagged and interfering MSs.

*Case* $m \neq n$: The inter-user interference expectation for $m \neq n$ can be expressed as

$$\begin{aligned}
&\mathbb{E}\left\{g_{mk'}g_{nk'}^{*}\varpi_{umk}\varpi_{unk}^{*}\right\}\\
&\quad=\mathbb{E}\left\{g_{mk'}\hat{g}_{mk}^{*}\right\}\mathbb{E}\left\{g_{nk'}^{*}\hat{g}_{nk}\right\}\\
&\quad=\begin{cases}\dfrac{\beta_{mk}\beta_{nk}}{\beta_{mk'}\beta_{nk'}}\mathbb{E}\left\{g_{mk'}\hat{g}_{mk'}^{*}\right\}\mathbb{E}\left\{g_{nk'}^{*}\hat{g}_{nk'}\right\}, & \boldsymbol{\varphi}_{k}=\boldsymbol{\varphi}_{k'}\\ \mathbb{E}\left\{g_{mk'}\right\}\mathbb{E}\left\{\hat{g}_{mk}^{*}\right\}\mathbb{E}\left\{g_{nk'}^{*}\right\}\mathbb{E}\left\{\hat{g}_{nk}\right\}, & \boldsymbol{\varphi}_{k}\neq\boldsymbol{\varphi}_{k'}\end{cases}\\
&\quad=\begin{cases}\dfrac{\beta_{mk}\beta_{nk}}{\beta_{mk'}\beta_{nk'}}\varrho_{mk'}\varrho_{nk'}, & \boldsymbol{\varphi}_{k}=\boldsymbol{\varphi}_{k'}\\ 0, & \boldsymbol{\varphi}_{k}\neq\boldsymbol{\varphi}_{k'}\end{cases}\\
&\quad=\frac{\beta_{mk}\beta_{nk}}{\beta_{mk'}\beta_{nk'}}\varrho_{mk'}\varrho_{nk'}\left|\boldsymbol{\varphi}_{k'}^{H}\boldsymbol{\varphi}_{k}\right|^{2},\quad (m\neq n)
\end{aligned}\tag{87}$$

for the UL, and

$$\begin{aligned}
&\mathbb{E}\left\{g_{mk}g_{nk}^{*}\varpi_{dmk'}\varpi_{dnk'}^{*}\right\}\\
&\quad=\mathbb{E}\left\{g_{mk}\frac{\hat{g}_{mk'}^{*}}{|\hat{g}_{mk'}|}\right\}\mathbb{E}\left\{g_{nk}^{*}\frac{\hat{g}_{nk'}}{|\hat{g}_{nk'}|}\right\}\\
&\quad=\begin{cases}\mathbb{E}\left\{g_{mk}\dfrac{\hat{g}_{mk}^{*}}{|\hat{g}_{mk}|}\right\}\mathbb{E}\left\{g_{nk}^{*}\dfrac{\hat{g}_{nk}}{|\hat{g}_{nk}|}\right\}, & \boldsymbol{\varphi}_{k}=\boldsymbol{\varphi}_{k'}\\ \mathbb{E}\left\{g_{mk}\right\}\mathbb{E}\left\{\dfrac{\hat{g}_{mk'}^{*}}{|\hat{g}_{mk'}|}\right\}\mathbb{E}\left\{g_{nk}^{*}\right\}\mathbb{E}\left\{\dfrac{\hat{g}_{nk'}}{|\hat{g}_{nk'}|}\right\}, & \boldsymbol{\varphi}_{k}\neq\boldsymbol{\varphi}_{k'}\end{cases}\\
&\quad=\begin{cases}\dfrac{\pi}{4}\sqrt{\varrho_{mk}\varrho_{nk}}, & \boldsymbol{\varphi}_{k}=\boldsymbol{\varphi}_{k'}\\ 0, & \boldsymbol{\varphi}_{k}\neq\boldsymbol{\varphi}_{k'}\end{cases}\\
&\quad=\frac{\pi}{4}\sqrt{\varrho_{mk}\varrho_{nk}}\left|\boldsymbol{\varphi}_{k'}^{H}\boldsymbol{\varphi}_{k}\right|^{2},\quad (m\neq n),
\end{aligned}\tag{88}$$

for the DL.

## APPENDIX B
## CALCULATION OF QUANTIZATION NOISE DISTORTION

### A. UL CASE

The quantization noise sample in the UL segment is given by (26) and, hence,

$$\begin{aligned}
&\mathbb{E}\left\{\left|\mathrm{QN}_{uk}\right|^{2}\right\}\\
&\quad=\begin{cases}\displaystyle\sum_{m=1}^{M}\sum_{n=1}^{M}\mathbb{E}\left\{\varpi_{umk}\varpi_{unk}^{*}\right\}\mathbb{E}\left\{\tilde{q}_{um}\tilde{q}_{un}^{*}\right\}, & \text{BCU}\\ \displaystyle\sum_{m=1}^{M}\sum_{n=1}^{M}\mathbb{E}\left\{\tilde{q}_{umk}\tilde{q}_{unk}^{*}\right\}, & \text{BAP.}\end{cases}
\end{aligned}\tag{89}$$

The inputs of the quantizers at different APs are correlated, since they arise from quantization of the same data signals. However, using the linear quantization model described in subsection II-B (see also [30]), the covariance matrix of the quantization distortion can be approximated as a diagonal matrix (see (8) in subsection II-B). This implies that the quantization distortion across APs can be reasonably approximated as uncorrelated [23, Appendix A]. Using this approximation in (89) then

$$\begin{aligned}
&\mathbb{E}\left\{\left|\mathrm{QN}_{uk}\right|^{2}\right\}\\
&\quad\approx\begin{cases}\displaystyle\sum_{m=1}^{M}\mathbb{E}\left\{\left|\hat{g}_{mk}\right|^{2}\right\}\mathbb{E}\left\{\left|\tilde{q}_{um}\right|^{2}\right\}, & \text{BCU}\\ \displaystyle\sum_{m=1}^{M}\mathbb{E}\left\{\left|\tilde{q}_{umk}\right|^{2}\right\}, & \text{BAP,}\end{cases}\\
&\quad=\begin{cases}\displaystyle\sum_{m=1}^{M}\varrho_{mk}\alpha_{um}\left(1-\alpha_{um}\right)\mathbb{E}\left\{\left|r_{um}\right|^{2}\right\}, & \text{BCU}\\ \displaystyle\sum_{m=1}^{M}\alpha_{umk}\left(1-\alpha_{umk}\right)\mathbb{E}\left\{\left|\hat{g}_{mk}^{*}r_{um}\right|^{2}\right\}, & \text{BAP,}\end{cases}
\end{aligned}\tag{90}$$

with

$$\mathbb{E}\left\{\left|r_{um}\right|^{2}\right\}=\sum_{k'=1}^{K}\nu_{k'}\beta_{mk'}+\sigma_{u}^{2},\tag{91}$$

and

$$\begin{aligned}
\mathbb{E}\left\{\left|\hat{g}_{mk}^{*}r_{um}\right|^{2}\right\}&=\sum_{k'=1}^{K}\nu_{k'}\beta_{mk'}\varrho_{mk}\\
&\quad+\sum_{k'=1}^{K}\nu_{k'}\frac{\beta_{mk}^{2}}{\beta_{mk'}^{2}}\varrho_{mk'}^{2}\left|\boldsymbol{\varphi}_{k'}^{H}\boldsymbol{\varphi}_{k}\right|^{2}+\varrho_{mk}\sigma_{u}^{2}.
\end{aligned}\tag{92}$$

### B. DL CASE

In the DL segment, the quantization noise sample at the $k$th MS is specified by (39) and, consequently, using again

the additive quantization noise model, its variance can be approximated as

$$\mathbb{E}\left\{\left|\mathrm{QN}_{dk}\right|^2\right\} \approx \begin{cases} \sum_{m=1}^{M} \beta_{mk}\mathbb{E}\left\{\left|\tilde{q}_{dm}\right|^2\right\}, & \text{BCU} \\ \sum_{m=1}^{M} \beta_{mk} \sum_{k'=1}^{K} \mathbb{E}\left\{\left|\tilde{q}_{dmk'}\right|^2\right\}, & \text{BAP} \end{cases} = \sum_{m=1}^{M} \beta_{mk} \sum_{k'=1}^{K} \tilde{\alpha}_{dmk'}\left(1-\tilde{\alpha}_{dmk'}\right)\eta_{mk'}. \quad (93)$$

## REFERENCES


[1] 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.

[2] T. Marzetta, E. Larsson, H. Yang, and H. Ngo, *Fundamentals of Massive MIMO* (Fundamentals of Massive MIMO). Cambridge, U.K.: Cambridge Univ. Press, 2016.

[3] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[4] 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.

[5] W. Choi and J. 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.

[6] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela, "Advances in smart antennas–network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Commun.*, vol. 13, no. 4, pp. 56–61, Aug. 2006.

[7] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[8] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[9] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile Networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

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

[11] S. Govindasamy and I. Bergel, "Uplink performance of multi-antenna cellular networks with co-operative base stations and user-centric clustering," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2703–2717, Apr. 2018.

[12] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2015, pp. 201–205.

[13] F. Riera-Palou, G. Femenias, A. G. Armada, and A. Perez-Neira, "Clustered cell-free massive MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[14] 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.

[15] L. D. Nguyen, T. Q. Duong, H. Q. Ngo, and K. Tourki, "Energy efficiency in cell-free massive MIMO with zero-forcing precoding design," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1871–1874, Aug. 2017.

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

[17] 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.

[18] 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.

[19] J. Kang, O. Simeone, J. Kang, and S. S. Shitz, "Joint signal and channel state information compression for the backhaul of uplink network MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1555–1567, Mar. 2014.

[20] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Layered downlink precoding for C-RAN systems with full dimensional MIMO," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2170–2182, Mar. 2017.

[21] P. Marsch and G. Fettweis, "On multicell cooperative transmission in backhaul-constrained cellular systems," *Ann. Telecommun. Annales des télécommunications*, vol. 63, nos. 5–6, pp. 253–269, Jun. 2008.

[22] P. Patil, B. Dai, and W. Yu, "Hybrid data-sharing and compression strategy for downlink cloud radio access network," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5370–5384, Nov. 2018.

[23] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, M. Debbah, and P. Xiao, "Max–min rate of cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6796–6815, Oct. 2019.

[24] M. N. Boroujerdi, A. Abbasfar, and M. Ghanbari, "Cell free massive MIMO with limited capacity fronthaul," *Wireless Pers. Commun.*, vol. 104, no. 2, pp. 633–648, Oct. 2018.

[25] Y. Zhang, M. Zhou, X. Qiao, H. Cao, and L. Yang, "On the performance of cell-free massive MIMO with low-resolution ADCs," *IEEE Access*, vol. 7, pp. 117968–117977, 2019.

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

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

[28] Z. Chen and E. Bjornson, "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.

[29] G. Interdonato, H. Q. Ngo, E. G. Larsson, and P. Frenger, "On the performance of cell-free massive MIMO with short-term power constraints," in *Proc. IEEE 21st Int. Workshop Comput. Aided Modelimg Design Commun. Links Netw. (CAMAD)*, Oct. 2016, pp. 225–230.

[30] A. Mezghani and J. A. Nossek, "Capacity lower bound of MIMO channels with output quantization and correlated noise," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Cambridge, MA, USA, Jul. 2012, pp. 1–5.

[31] J. Bussgang, *Crosscorrelation Functions of Amplitude-Distorted Gaussian Signals*. Cambridge, MA, USA: MIT Press, Mar. 1952.

[32] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Robust predictive quantization: Analysis and design via convex optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 618–632, Dec. 2007.

[33] E. Bjornson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[34] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inf. Theory*, vol. IT-6, no. 1, pp. 7–12, Mar. 1960.

[35] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, vol. 159. Springer, 2012.

[36] J. Mo and R. W. Heath, Jr., "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5498–5512, Oct. 2015.

[37] L. Fan, S. Jin, C.-K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2186–2189, Dec. 2015.

[38] J. Zhang, L. Dai, S. Sun, and Z. Wang, "On the spectral efficiency of massive MIMO systems with low-resolution ADCs," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 842–845, May 2016.

[39] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, May 2016.

[40] W. Tan, S. Jin, C.-K. Wen, and Y. Jing, "Spectral efficiency of mixed-ADC receivers for massive MIMO systems," *IEEE Access*, vol. 4, pp. 7841–7846, 2016.

[41] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath, Jr., "Hybrid architectures with few-bit ADC receivers: Achievable rates and energy-rate tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2274–2287, Apr. 2017.

[42] K. Roth and J. A. Nossek, "Achievable rate and energy efficiency of hybrid and digital beamforming receivers with low resolution ADC," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2056–2068, Sep. 2017.
[43] L. N. Ribeiro, S. Schwarz, M. Rupp, and A. L. F. de Almeida, "Energy efficiency of mmWave massive MIMO precoding with low-resolution DACs," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 298–312, May 2018.
[44] X. Hu, C. Zhong, X. Chen, W. Xu, H. Lin, and Z. Zhang, "Cell-free massive MIMO systems with low resolution ADCs," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6844–6857, Oct. 2019.
[45] J. Xu, W. Xu, H. Zhang, G. Y. Li, and X. You, "Performance analysis of multi-cell millimeter-wave massive MIMO networks with low-precision ADCs," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 302–317, Jan. 2019.
[46] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO—5G system," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 905–923, 2nd Quart., 2016.
[47] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao, "Energy efficiency of the cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 971–987, Dec. 2019.

**GUILLEM FEMENIAS** (Senior Member, IEEE) received the degree in telecommunication engineering and the Ph.D. degree in electrical engineering from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1987 and 1991, respectively. From 1987 to 1994, he worked as a Researcher with the UPC, where he became an Associate Professor, in 1992. In 1995, he joined the Department of Mathematics and Informatics, University of the Balearic Islands (UIB), Mallorca, Spain, where he became a Full Professor, in 2010. He is currently leading the Mobile Communications Group, UIB. In the past, he was also involved in several European projects such as ATDMA, CODIT, and COST. On his research topics, he has published more than 100 journal articles and conference papers, as well as some book chapters. He has been the Project Manager of projects such as ARAMIS, DREAMS, DARWIN, MARIMBA, COSMOS, ELISA, and TERESA, all of them funded by the Spanish and Balearic Islands Governments. His current research interests and activities span the fields of digital communications theory and wireless communication systems, with a particular emphasis on radio resource management strategies applied to 5G and 6G wireless networks. He was the recipient of the Best Paper Awards from the IFIP International Conference on Personal Wireless Communications, in 2007, and the IEEE Vehicular Technology Conference-Spring, in 2009. He has served as a technical program committee member of various IEEE conferences, the Publications Chair of the IEEE 69th Vehicular Technology Conference (VTC-Spring), in 2009, and a Local Organizing Committee Member of the IEEE Statistical Signal Processing (SSP), in 2016.

**FELIP RIERA-PALOU** (Senior Member, IEEE) received the B.S. and M.S. degrees in computer engineering from the University of the Balearic Islands (UIB), Mallorca, Spain, in 1997, the M.Sc. and Ph.D. degrees in communication engineering from the University of Bradford, U.K., in 1998 and 2002, respectively, and the M.Sc. degree in statistics from The University of Sheffield, U.K., in 2006. From May 2002 to March 2005, he was with Philips Research Laboratories, Eindhoven, The Netherlands, first as a Marie Curie Postdoctoral Fellow (European Union) and later as a Member of the Technical Staff. While at Philips, he worked on research programs related to wideband speech/audio compression and speech enhancement for mobile telephony. From April 2005 to December 2009, he was a Research Associate (Ramon y Cajal Program, Spanish Ministry of Science) with the Mobile Communications Group, Department of Mathematics and Informatics, UIB, where he has been an Associate Research Professor (I3 Program, Spanish Ministry of Education), since January 2010. His current research interests include signal processing and wireless communications.