Uplink Performance Optimization of Limited-Capacity Radio Stripes

Ioannis Chiotis, *Graduate Student Member, IEEE*, and Aris L. Moustakas, *Senior Member, IEEE*

***Abstract*— Cell-free (CF) massive multiple-input multiple-output (mMIMO) is a network architecture beyond fifth generation (B5G), which has the potential to deliver significantly higher spectral efficiency (SE) and energy conservation, when compared to the traditional cellular MIMO layout. Radio stripes form a particular realization of CF mMIMO topologies, which at a relatively modest deployment cost, promise to distribute part of the computational load of the centralized processing unit (CPU), while maintaining the same performance. However, the limited-capacity fronthaul (FH) network effect has not yet been adequately studied in this context. In this paper, we develop an uplink sequential processing algorithm that is optimal in the sense of locally minimizing the mean-squared error (MSE) at every antenna processing unit (APU). The performance is further enhanced by applying an effective Compare-and-Forward (CnF) strategy or by minimizing the compression error covariance trace, given the fronthaul capacity constraint. Additionally, we study the case where the radio stripe arrangement uses a distributed setup of access points (APs), aiming at minimizing path attenuation even more effectively. Based on analytical equations and simulation results, we conclude that throughput is maximized when the classic radio stripe setup is combined with the proposed algorithm and the CnF technique.**

***Index Terms*— B5G, user-centric radio stripe, cell-free massive MIMO, limited-capacity fronthaul, spectral efficiency, optimal sequential processing, distributed processing.**

I. Introduction

MASSIVE multiple-input multiple-output (mMIMO) is one of the most promising architectures for high data rates [3], [4], [5], [6]. Nevertheless, conventional mMIMO can be impeded by strong signal fluctuations and poor performance of the users at the cell edges. One innovative idea that tackles these issues is the so-called Cell-Free (CF) mMIMO [7], [8], [9] network topology. This refers to a distributed mMIMO setup that is capable of implementing coherent service to all the nearby user equipment (UEs), yet avoiding the creation of cell boundaries, a key factor that provides additional macro-diversity, reduces path loss and overcomes inter-cell interference limitations.

The original CF mMIMO layout consists of a centralized processing unit (CPU) that is directly connected to multiple access points (APs), via dedicated fronthaul (FH) connections, essentially through a star-like topology that can jointly serve a smaller number of distributed UEs [8], [9], [10]. To achieve that, all APs act coherently and serve all the UEs of the network in the same frequency-time frame via time-division duplex (TDD) operation. Although this architecture, especially when combined with minimum mean-squared error (MMSE) processing [9], leads to significantly higher spectral efficiency (SE) [8] and energy efficiency (EE) [11], [12] when compared to collocated mMIMO, it is noticeably more costly to realize, since it demands significant capital expenditures to lay a dense network of long wires that will connect each individual AP to the CPU. This factor, coupled with the increased FH signaling and computational complexity, greatly confines the scalability of these systems, making their practical implementation quite challenging.

The need for decentralization and scalability directed researchers to more distributed signal processing approaches [9], [13], [14], [15], [16], [17], [18], [19], [20], that can be applied to CF mMIMO architectures and thus can alleviate most of the aforementioned issues. Another promising direction is that of the so-called radio stripes [1], [2], [7], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. A radio stripe is actually a piece of flexible material (e.g. fibre or copper wire) [23] that provides sequential connectivity, power supply and data transfer to many antenna processing units (APUs). Each APU integrates a small number of antennas (conventional $\lambda/2$ dipoles [23] or more advanced, e.g. uniplanar antennas that employ spoof surface plasmon polariton [31]), henceforth referred to as APs, and is responsible for fast processing the data they collect. This serial interconnection of the APUs essentially allows for sequential signal processing, a key feature that can offer the benefits of a classic CF mMIMO, yet providing low infrastructure cost, reduced FH signaling and a decentralized architecture.

A. Motivation and Related Work

In most cases, the computational complexity and the increased FH signaling of the centralized CF mMIMO systems

Manuscript received 9 August 2023; revised 17 December 2023 and 16 February 2024; accepted 6 April 2024. Date of publication 29 April 2024; date of current version 12 September 2024. This work was supported in part by the National Recovery and Resilience Plan "Greece 2.0" funded by European Union under the NextGenerationEU Program under Project MIS 5154714. An earlier version of this paper was presented in part at IEEE MeditCom 2022 [DOI: 10.1109/MeditCom55741.2022.9928628] and in part at IEEE ISNCC 2023 [DOI: 10.1109/ISNCC58260.2023.10323670]. The associate editor coordinating the review of this article and approving it for publication was X. Tao. *(Corresponding author: Ioannis Chiotis.)*

Ioannis Chiotis is with the Department of Physics, National and Kapodistrian University of Athens, 15784 Athens, Greece (e-mail: ioachiotis@phys.uoa.gr).

Aris L. Moustakas is with the Department of Physics, National and Kapodistrian University of Athens, 15784 Athens, Greece, and also with the Archimedes/Athena Research Unit, 15125 Athens, Greece.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2024.3392178.

Digital Object Identifier 10.1109/TWC.2024.3392178

escalate rapidly with the number of UEs, thus producing practical issues (i.e. limited scalability). The initial idea to mitigate these drawbacks was to equip the APs with some computational capabilities, in order to perform part of the overall signal processing [9], [14], [20], [32], [33]. Another well-known concept towards that direction is the so-called dynamic cooperation clustering (DCC) [34], [35], which essentially allows each UE to be served only from a subset of APs, according to a specif criterion. That UE-AP association problem is the subject of many researches, as it determines the counterbalancing relation between the network performance [7], [36] and the required FH signaling, computational complexity [14], [15], [17], [18], [34], [37] and power consumption [11]. Nevertheless, state-of-the-art studies have shown that there are methods capable of achieving near-optimal rates, while sustaining scalability [13], [14], [38] and low-capacity requirements [13].

However, despite the above mentioned benefits that the advanced CF mMIMO approaches can offer, the need for extensive FH link installations still remains. Daisy chain architectures, i.e. radio stripes [1], [2], [7], [21], [22], [23], [24], [29], form a versatile and cost-effective alternative for deploying the CF mMIMO FH network in dense environments, while thanks to their structure, they also allow for decentralized processing. The simplest algorithm for use in radio stripes was observed in [7], where authors demonstrated a sequential method of performing the maximum ratio (MR) combining/precoding. Nevertheless, the first uplink algorithm that truly allowed the serially connected APUs to cooperate, was the normalized linear MMSE (N-LMMSE) [1], [22]. That scheme enabled every APU to improve each UE's soft estimation, by combining information originated from both its assigned APs and the preceding APU. By applying this methodology, it was possible to gradually mitigate multi-user interference, thus making radio stripes more competitive against conventional setups that use maximum ratio combining (MRC) [22]. Another innovative algorithm was the regularized zero-forcing (C-RZF) [39], which was capable of achieving almost optimal results, while keeping latency and FH signaling at low levels. Ultimately, the optimum performance of radio stripes was achieved by the optimal sequential linear processing (OSLP) algorithm [21], which performed equally to the level 4 (L4) centralized MMSE [9]. Nevertheless, recent studies [28] have shown that by employing the OSLP, or even suboptimal MR combiners, jointly with an access point selection (APS) strategy, can actually lead to a higher network throughput. However, that is done by solving a compute-intensive bi-objective [28] optimization problem which requires knowledge over the instantaneous signal-to-interference-plus-noise ratio (SINR) of each user, a metric that has to be calculated at the CPU, thus distorting the sequential procedure. Finally, a near-optimal [as per the bit-error-rate (BER) metric] sequential uplink algorithm, based on the Gaussian message detection (GMD) [40] method and which ensures a constant FH signaling and a low complexity that scales linearly with the number of UEs, was examined in [29].

Besides the focus on strategies that only intend to achieve higher UE rates, radio stripes extend to other research ares as well. For instance, in [30], authors model an uplink transmission channel for line-of-sight (LOS) communications, mainly at the millimeter wave (mmWave) band, whereas in [26], multiple phase-synchronized radio stripes are used for sensing reasons, namely to jointly locate and synchronize terminal devices. Also, their applicability for efficient wireless energy transfer (WET) purposes is examined in [27].

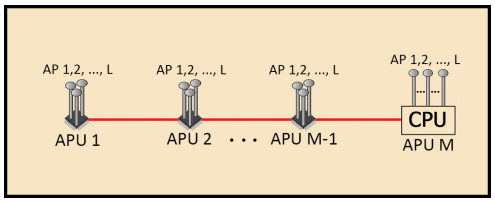
Despite the progress towards that direction, none of the above works takes into account the limited capacity that is inherent in every link between the APUs. Nevertheless, there are numerous papers that analyze the impact of finite capacity on both cellular [41], [42] and CF [43], [44], [45], [46], [47] mMIMO implementations. For instance, in [43], authors examined the application of a max-min optimization problem under the presence of uniformly distributed compression noise [48, Ch. 2]. Later, by employing the Bussgang's decomposition, they calculated the optimal step size Δ_{opt} [46] that maximizes the signal-to-distortion-noise ratio (SDNR). This solution was then used in [47], for further investigation towards the throughput and the energy efficiency improvement of the system under study. However, their approach included perfect hardware, scalar compressions and equal bit allocations for each transmission. The impact of finite-capacity on the SE and EE of a CF mMIMO setup, under the assumption of hardware impairments, vector quantization and Gaussian compression noise, was seen in [44].

Finally, the analysis of the effects of the limited-capacity constraint on a radio stripe system has recently been studied in a small number of works. In particular, its impact on the downlink performance of a radio stripe was investigated in [24] and [49], while, in [1] and [2], the current authors dealt with the same effects for the uplink scenario. Specifically, in [1], we introduced the Compare-and-Forward (CnF) strategy and applied it to the suboptimal N-LMMSE algorithm [22], whereas in [2], we derived the generalization of the OSLP algorithm [21], also considering all the finite-capacity links between the APUs, and compared it with the N-LMMSE, with and without the application of CnF.

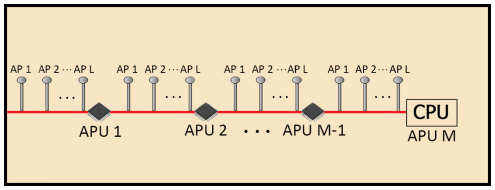
B. Contributions

In this paper we build upon [1] and [2] to analyze the effects that a finite-capacity FH link has on a radio stripe network and to propose ways to mitigate its impact. Specifically, we optimize the compression noise covariance matrix, taking into account the finite-capacity restriction. We also optimize the performance of the distributed radio stripe setup of Fig. 1b, and compare it to that of Fig. 1a. Furthermore, we discuss the complexity of the algorithms and show that the CnF strategy not only offers higher throughput, but also can significantly lower the average total latency. In conclusion, the main contributions of our work can be summarized as follows:

- By considering the existence of a limited-capacity radio stripe network, we develop an uplink sequential processing algorithm that includes the incurred compression



(a) Radio stripe with collocated APs.



(b) Radio stripe with fully distributed APs.

Fig. 1. Radio stripe models.

noise, appearing at each APU, and which is proved to be optimal in the sense of minimizing the produced MSE.

- In addition to the standard radio stripe setup, we also propose and optimize the performance of a novel arrangement (Fig. 1b) that embeds APUs with a distributed number of APs. This layout not only allows for a larger number of independent antennas per APU, but also brings these antennas closer to the UE terminals.
- We introduce a novel CnF strategy (Algorithm 1) which, by assigning each UE to a specific dynamic cluster of APUs and thus avoiding unnecessary signal compressions, can ensure a higher throughput and a reduced service latency to the majority of the users in the network.
- We derive the optimal quantization noise covariance matrix (Algorithm 2), which subject to the fronthaul link capacity constraints minimizes its trace, thus further suppressing the MSE locally at each APU.
- We discuss about the complexity and latency issues of the above algorithms and derive achievable SE expressions for each individual case. Finally, using this metric, we numerically evaluate the performance of every analyzed algorithm and radio stripe setup.

C. Paper Outline

In Section II, we briefly present the system model for the uplink case of a classic CF mMIMO, including both pilot and payload signal transmissions. We also derive some basic distortion-rate expressions that will be used later on. Section III includes the derivation of the optimal sequential processing algorithm (OSPA) as well as closed-form expressions for the uplink per-user SINR. In Section IV, we further augment our algorithm by proposing a CnF strategy that enables dynamic cooperation clustering, thus leading to a *user-centric* system. An alternative quantization error suppression scheme is investigated in Section V, while in Section VI we discuss about the complexity and latency issues of the derived algorithms and present the numerical results. Ultimately, the paper is concluded in Section VII.

D. Notations

The superscripts $(.)^*$, $(.)^\dagger$, $(.)^T$ and $(.)^{-1}$ stand for the conjugate, conjugate transpose, transpose and inverse, respectively. Notations $\mathbb{E}\{.\}$, tr(.), $|.|$, $\|.\|$ and $\triangleq$ denote the expected value, the trace of a matrix, the absolute value of a scalar, the l_2 norm of a vector and definitions, respectively. Boldface lowercase letters (e.g. $\mathbf{v}$) denote column vectors, boldface uppercase letters (e.g. $\mathbf{B}$) denote matrices and $\mathbf{I}_L$ denotes the $L \times L$ identity matrix. Circularly-symmetric variables that follow complex normal distribution with covariance matrix $\mathbf{R}$ and zero mean are represented as $\mathcal{CN}(\mathbf{0}, \mathbf{R})$, while block-diagonal matrices as $\text{diag}(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_K)$, with $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_K$ being square matrices placed on the diagonal block. Also, the set $\{1, 2, \ldots, M\}$ is denoted as [M]. Finally, we use standard notations for differential entropy, mutual information and complexity.

II. Radio Stripe Network Model

We envision a general radio stripe architecture, comprising of M APUs (the Mth is the CPU), each one connected with L single-antenna APs (either distributed along the stripe using part of the total capacity [1] or collocated on every APU [21], [22]). In the setup we consider K single-antenna UEs, where the channel between them and the L APs of each APU m, where $m \in [M]$, is given by $\mathbf{G}_m = [\mathbf{g}_{m1}, \mathbf{g}_{m2}, \ldots, \mathbf{g}_{mK}] \in \mathbb{C}^{L \times K}$. Elements $\mathbf{g}_{mk} \in \mathbb{C}^L$ are constant over a coherence interval τ_c and are drawn from an uncorrelated Rayleigh fading distribution as

$$\mathbf{g}_{mk} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{mk}) \tag{1}$$

where $\mathbf{R}_{mk} \in \mathbb{C}^{L \times L}$ is the diagonal covariance matrix with diagonal entries β_{mkl}, for $l = 1, 2, \ldots, L$, which is assumed to be known at the APUs. The average large-scale fading coefficient that describes the path loss and shadowing is denoted by $\beta_{mk} \triangleq \text{tr}(\mathbf{R}_{mk})/L$.

Henceforward the total communication bandwidth will be denoted as B and the coherence bandwidth of the channel as B_c. Thus there are B/B_c independent channel elements in frequency space. Similarly, the temporal coherence time, e.g. due to mobility, over which the channel can be assumed to be constant, will be designated as T_c. Hence, there are $\tau_c = B_c T_c$ symbols that can be transmitted over frequency-time space. From these, $\tau_p \leq \tau_c$ are allocated to channel training, while the rest $\tau_d = \tau_c - \tau_p$ to payload data.

A. Fronthaul Compression and Transmission

As seen in Fig. 1, the architectures under study include communications between APs and APUs, as well as among successive APUs via wired (non-wireless) FH links. These links may be through copper, such DSL-based systems, or optical-based systems. More specifically, in Fig. 1a, FH interconnections are more simple, as APs are collocated on each APU and thus pilot and payload signals do not need compression. Hence, all the available capacity of the radio stripe is dedicated for the transmission signals, in the form of quantized IQ samples, only among APUs. However, in Fig. 1b where APs are also distributed, we need to take into account the finite-capacity constraint between them and the APUs as well. In any case, the finite rate of these links necessitates the quantization to be at a level which can be transmitted via

the FH link without error. This effectively adds a quantization noise to the samples, which can then be associated with the FH link requirements using information theoretic expressions.

Let $\mathbf{x} \in \mathbb{C}^L$ be a random vector with zero mean and variance $\mathbb{E}\{\mathbf{x}\mathbf{x}^\dagger\} = \mathbf{P}$. Provided that the length L of sequence $\mathbf{x}$ is quite large, we will compress that sequence using the following test channel

$$\bar{\mathbf{x}} = \mathbf{x} + \mathbf{w} \tag{2}$$

where $\mathbf{w} \in \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$ represents the quantization noise that is assumed to be independent of the signal $\mathbf{x}$ [42], [44], [50]. Then, given the available rate $\mathcal{R}$ of the test channel in (2), we can relate it with the quantization error covariance matrix $\mathbf{\Sigma}$ as [50, Ch. 9]

$$\begin{aligned}\mathcal{R} &= I(\mathbf{x};\bar{\mathbf{x}}) \\ &= h(\bar{\mathbf{x}}) - h(\mathbf{x}+\mathbf{w}|\mathbf{x}) \\ &\overset{(a)}{=} h(\bar{\mathbf{x}}) - h(\mathbf{w}) \\ &\overset{(b)}{\leq} \log_2 \det\left(\mathbf{P}\mathbf{\Sigma}^{-1} + \mathbf{I}_L\right)\end{aligned} \tag{3}$$

where (a) derives from the independence between $\mathbf{x}$ and $\mathbf{w}$ and (b) from the *maximum differential entropy lemma* [51, Ch. 2]. Finally, notice that the application of the upper bound in (3) overestimates the required FH rate $\mathcal{R}$ needed in order to transfer the compressed signal over the link. Thus, given a specific link rate, a stronger compression noise is considered.

Consequently, if we assume that the total capacity of each FH link that interconnects neighboring APUs is C, we can divide it into two rates. The first rate, denoted by $C_{ap} = rC$, is associated with communications between the APs and their APU, while the second rate, denoted by $C_{apu} = (1-r)C$, regards communications between the APUs, where ratio[1] $0 \leq r < 1$ will eventually be optimized. For simplicity, we will assume that each AP uses equal rate C_{ap}/L for both pilot and payload signal transferring to its assigned APU. Thus, C_{ap} can be further split into C_{ap}^p and C_{ap}^d, so that $C_{ap} = C_{ap}^p + C_{ap}^d$, which represent the allocated rates for the transmission of pilot and payload signals from the APs to the APUs, respectively. Finally, to be more fair, these rates are set to be proportional of the interval they serve as $C_{ap}^p = \frac{\tau_p}{\tau_c} C_{ap}$ and $C_{ap}^d = \frac{\tau_d}{\tau_c} C_{ap}$.

B. Uplink Training Period

Let τ_p be the duration (in samples) of the uplink training period, where $\tau_p \leq \tau_c$. During that phase, all K UEs simultaneously transmit τ_p-length mutually orthogonal pilot sequences towards the APs. The pilot sequence of each UE k, where $k \in [K]$, is designated as $\boldsymbol{\phi}_k \in \mathbb{C}^{\tau_p}$, with $\|\boldsymbol{\phi}_k\|^2 = \tau_p$. Also, for simplicity reasons, we do not include the effects of pilot contamination ($K > \tau_p$), although, in that case, the derived equations in the current analysis would remain the same. Nonetheless, pilot contamination would decrease the achievable performance of each UE due to the increase of the channel estimation errors and the FH signaling between the APUs (stronger compression noise). For concreteness, we only consider the case where $K = \tau_p$.

The signal vector $\mathbf{y}_{p,m\tau} \in \mathbb{C}^L$ received at random channel use τ, where τ is within τ_p, by the L APs of every APU m can be written as follows

$$\mathbf{y}_{p,m\tau} = \mathbf{G}_m \boldsymbol{\psi}_\tau^T + \mathbf{n}_m \tag{4}$$

where $\boldsymbol{\psi}_\tau = [\sqrt{\rho_1}\phi_{\tau 1}, \sqrt{\rho_2}\phi_{\tau 2}, \ldots, \sqrt{\rho_K}\phi_{\tau K}] \in \mathbb{C}^{1\times K}$ is the pilot vector that contains the τth entry of all pilot signals $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K$ and $\mathbf{n}_m$ the receiver noise vector with independent and identically distributed (i.i.d.) $\mathcal{CN}(0, \sigma_n^2)$ entries. Then, all the L APs of each APU m compress their received signal according to (2). The instantaneous quantized signal $\bar{\mathbf{y}}_{p,m\tau}$ that finally reaches the mth APU is

$$\bar{\mathbf{y}}_{p,m\tau} = \mathbf{G}_m \boldsymbol{\psi}_\tau^T + \mathbf{n}_m + \mathbf{w}_m^p \tag{5}$$

with $\mathbf{w}_m^p \in \mathbb{C}^L \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_m^p)$ being the quantization error vector due to the imperfect pilot compression. Therefore, according to (3) and given that the available FH rate for the transmission of the instantaneous $\mathbf{y}_{p,m\tau}$ from the L APs (of each APU m) to the APU m is C_{ap}^p/τ_p, the covariance matrix $\mathbf{\Sigma}_m^p$ is calculated as

$$\begin{aligned}\frac{C_{ap}^p}{\tau_p} &= \frac{1}{\tau_c}\log_2 \det\left(\mathbb{E}\{\mathbf{y}_{p,m\tau}\mathbf{y}_{p,m\tau'}^\dagger\}\mathbf{\Sigma}_m^{p-1} + \mathbf{I}_L\right) \\ &= \frac{L}{\tau_c}\log_2\left(\frac{\sum_{i=1}^K \rho_i\beta_{mi} + \sigma_n^2}{p_m} + 1\right)\end{aligned} \tag{6}$$

where $\mathbf{\Sigma}_m^p = p_m\mathbf{I}_L$ and the expectation in (6) has been calculated over many coherence intervals τ_c. Hence, each entry of $\mathbf{\Sigma}_m^p$ can be expressed as

$$p_m = \frac{\sum_{i=1}^K \rho_i\beta_{mi} + \sigma_n^2}{2^{\frac{C_{ap}}{L}} - 1} \tag{7}$$

At the end of the uplink training period, the total quantized pilot matrix $\bar{\mathbf{Y}}_{p,m}$ received by each APU m is represented as

$$\bar{\mathbf{Y}}_{p,m} = \mathbf{G}_m\mathbf{\Phi}^T + \mathbf{N}_m + \mathbf{W}_m^p \tag{8}$$

where $\mathbf{\Phi} = [\boldsymbol{\psi}_1^\dagger, \boldsymbol{\psi}_2^\dagger, \ldots, \boldsymbol{\psi}_{\tau_p}^\dagger]^\dagger \in \mathbb{C}^{\tau_p\times K}$ is the total pilot matrix. Also, $\mathbf{N}_m \in \mathbb{C}^{L\times\tau_p}$ expresses the additive white Gaussian noise matrix and $\mathbf{W}_m^p \in \mathbb{C}^{L\times\tau_p}$ the pilot quantization error matrix at the mth APU, both assumed to contain i.i.d. complex Gaussian elements with zero mean and variance σ_n^2 and p_m, respectively. Using (8), the MMSE channel estimate $\hat{\mathbf{g}}_{mk} \in \mathbb{C}^L$ is given as [5, Sec. 3]

$$\hat{\mathbf{g}}_{mk} = \sqrt{\rho_k\tau_p}\mathbf{R}_{mk}\mathbf{\Psi}_{mk}^{-1}\check{\mathbf{y}}_{p,mk} \tag{9}$$

where $\check{\mathbf{y}}_{p,mk}$ and $\mathbf{\Psi}_{mk}$ are given from

$$\check{\mathbf{y}}_{p,mk} = \bar{\mathbf{Y}}_{p,m}\frac{\boldsymbol{\phi}_k^*}{\sqrt{\tau_p}} = \sqrt{\rho_k\tau_p}\mathbf{g}_{mk} + \mathbf{N}_m\frac{\boldsymbol{\phi}_k^*}{\sqrt{\tau_p}} + \mathbf{W}_m^p\frac{\boldsymbol{\phi}_k^*}{\sqrt{\tau_p}} \tag{10}$$

$$\mathbf{\Psi}_{mk} = \mathbb{E}\{\check{\mathbf{y}}_{p,mk}\check{\mathbf{y}}_{p,mk}^\dagger\} = \rho_k\tau_p\mathbf{R}_{mk} + (\sigma_n^2 + p_m)\mathbf{I}_L \tag{11}$$

The error that results from the imperfect channel estimation is denoted as $\tilde{\mathbf{g}}_{mk} = \mathbf{g}_{mk} - \hat{\mathbf{g}}_{mk}$ and it is independent from the channel estimation $\hat{\mathbf{g}}_{mk}$, with $\hat{\mathbf{g}}_{mk} \sim \mathcal{CN}(\mathbf{0}, \rho_k\tau_p\mathbf{R}_{mk}\mathbf{\Psi}_{mk}^{-1}\mathbf{R}_{mk})$ and $\tilde{\mathbf{g}}_{mk} \sim \mathcal{CN}(\mathbf{0}, \tilde{\mathbf{R}}_{mk})$. The covariance matrix $\tilde{\mathbf{R}}_{mk}$ is defined as

$$\tilde{\mathbf{R}}_{mk} = \mathbb{E}\{\tilde{\mathbf{g}}_{mk}\tilde{\mathbf{g}}_{mk}^\dagger\} = \mathbf{R}_{mk} - \rho_k\tau_p\mathbf{R}_{mk}\mathbf{\Psi}_{mk}^{-1}\mathbf{R}_{mk} \tag{12}$$

[1]That ratio equals zero only in case where APs are collocated on their respective APUs. In every other scenario, r varies between values zero and one.

C. Uplink Payload Period

During the uplink payload period, all K users simultaneously transmit their data towards the APs for a duration of $\tau_d = \tau_c - \tau_p$ samples. For each channel use τ within τ_d, the instantaneous compressed signal $\bar{\mathbf{y}}_m$ received by the mth APU is

$$\bar{\mathbf{y}}_m = \mathbf{G}_m \mathbf{q} + \mathbf{n}_m + \mathbf{w}_m^d \tag{13}$$

where $\mathbf{q} = [q_1, q_2, \ldots, q_K]^T \in \mathbb{C}^K \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q})$ is the transmitted message vector, with the covariance $\mathbf{Q} = \text{diag}(\rho_1, \rho_2, \ldots, \rho_K)$ being the transmitted power matrix of all UEs. Also, $\mathbf{w}_m^d$ is the quantization error vector due to the imperfect compression of $\mathbf{y}_m$, assumed to contain i.i.d. $\mathcal{CN}(0, d_m)$ entries. In a similar way to (6), (7), it derives that d_m is equal to p_m.

Furthermore, considering that $\mathbf{G}_m = \hat{\mathbf{G}}_m + \tilde{\mathbf{G}}_m$, with $\hat{\mathbf{G}}_m = [\hat{\mathbf{g}}_{m1}, \hat{\mathbf{g}}_{m2}, \ldots, \hat{\mathbf{g}}_{mK}]$ being the channel estimation matrix and $\tilde{\mathbf{G}}_m = [\tilde{\mathbf{g}}_{m1}, \tilde{\mathbf{g}}_{m2}, \ldots, \tilde{\mathbf{g}}_{mK}]$ the channel error matrix, (13) can be rewritten as

$$\bar{\mathbf{y}}_m = \hat{\mathbf{G}}_m \mathbf{q} + \boldsymbol{\xi}_m \tag{14}$$

where the total noise vector $\boldsymbol{\xi}_m = \tilde{\mathbf{G}}_m \mathbf{q} + \mathbf{n}_m + \mathbf{w}_m^d$ is distributed as $\mathcal{CN}(\mathbf{0}, \boldsymbol{\Omega}_m)$, with $\boldsymbol{\Omega}_m$ to be given as

$$\boldsymbol{\Omega}_m = \sum_{i=1}^{K} \rho_i \tilde{\mathbf{R}}_{mi} + (\sigma_n^2 + d_m)\mathbf{I}_L \tag{15}$$

III. Sequential Processing Analysis

As discussed in the previous section, radio stripe geometries of Fig. 1 require that the estimated signal vector, at a given APU, is transmitted via a finite-capacity link to the next APU in the line. This next APU may then combine the received vector with the payload signals which has collected from the L APs associated with it. The final outcome may then be transmitted further, via the next finite-capacity link, to the next APU. That process keeps on until the signal estimation vector finally reaches the CPU. Thus, we can see that at each step, the estimated signal vector gets degraded due to the addition of quantization noise from the finite-capacity links, yet at the same time, it also gets improved due to the addition of new information.

In the present section we propose an optimal uplink processing algorithm, which combines, at each APU m, the noisy soft estimation vector from APU $m-1$ with the payload signal that has obtained from its L APs. The algorithm is a generalization of [21], taking also into account all the finite-capacity links and it is *optimal* in the sense of minimizing the MSE at each APU. However, and in contrast to [21], our case involves the addition of the compression noise right *after* the combining process, thus leading to an inferior performance than this of a centralized processing system with infinite-capacity links (equivalent to the OSLP algorithm in [21]). Hence, performance enhancement is not guaranteed for every user at each step of the proposed algorithm, an affect that we manage to counterbalance with the introduction of two novel strategies in Sections IV and V.

To describe the algorithm, we must first define a number of quantities. Assume that $\hat{\mathbf{s}}_m \in \mathbb{C}^K$ is the signal soft estimation vector at the mth APU, where $m \in [M]$, prior to its compression.[2] After its compression and transmission via the mth finite-capacity link, it is denoted with $\bar{\mathbf{s}}_m \in \mathbb{C}^K$. These two vectors are associated, according to (2), as follows

$$\bar{\mathbf{s}}_m = \hat{\mathbf{s}}_m + \mathbf{w}_m \tag{16}$$

where $\mathbf{w}_m \in \mathbb{C}^K$ is the quantization noise vector that is inserted to $\hat{\mathbf{s}}_m$ due to the imperfect compression. Then, considering a separate quantizer for each entry $\hat{s}_{m,k}$ of $\hat{\mathbf{s}}_m$ [41], [42], [44], the variance of each element $w_{m,k} \sim \mathcal{CN}(0, \mathbb{E}\{|w_{m,k}|^2\})$ within the vector $\mathbf{w}_m$ can be calculated, based on (3), as

$$\mathbb{E}\{|w_{m,k}|^2\} = \epsilon \mathbb{E}\{|\hat{s}_{m,k}|^2\}, \quad \forall k \in [K] \tag{17}$$

where $\epsilon = (2^{\frac{C_{apu}\tau_c}{N}} - 1)^{-1}$ and N is the total number of complex scalars exchanged between the APUs.[3]

Remark 1: Since we have assumed that each element within $\hat{\mathbf{s}}_m$ is compressed via a different quantizer, one can elicit that quantization noises are independent with each other, a condition that serves two purposes. The first purpose is the simplicity of the calculations, while the second is that this design enables us to apply the CnF strategy without altering the compression error variances $\mathbb{E}\{|w_{m,k}|^2\}$ (see Sections IV and V). However, from an information theoretic point of view, we know that since the correlation between the elements of $\hat{\mathbf{s}}_m$ is neglected, this design leads to suboptimal results. Even so, optimality could still be achieved by applying the reverse water-filling method, for a high-resolution regime [50, Ch. 10].

Then, at each APU m, we can express $\bar{\mathbf{s}}_m$ as a linear combination of all the received signals $\bar{\mathbf{y}}_{m'}$ and all the quantization error vectors $\mathbf{w}_{m'}$, which have been inserted at each finite-capacity link m', where $m' \leq m$. Thus, $\bar{\mathbf{s}}_m$ can be written as follows

$$\bar{\mathbf{s}}_m = \overline{\mathbf{B}}_m^0 \bar{\boldsymbol{\upsilon}}_m + \overline{\boldsymbol{\Gamma}}_m^0 \boldsymbol{\omega}_m \tag{18}$$

where $\boldsymbol{\omega}_m = [\mathbf{w}_1^\dagger, \mathbf{w}_2^\dagger, \ldots, \mathbf{w}_m^\dagger]^\dagger \in \mathbb{C}^{mK}$ and $\bar{\boldsymbol{\upsilon}}_m = [\bar{\mathbf{y}}_1^\dagger, \bar{\mathbf{y}}_2^\dagger, \ldots, \bar{\mathbf{y}}_m^\dagger]^\dagger \in \mathbb{C}^{mL}$, which can also be written as

$$\bar{\boldsymbol{\upsilon}}_m = \hat{\mathbf{H}}_m \mathbf{q} + \boldsymbol{\zeta}_m \tag{19}$$

where $\hat{\mathbf{H}}_m = [\hat{\mathbf{G}}_1^\dagger, \hat{\mathbf{G}}_2^\dagger, \ldots, \hat{\mathbf{G}}_m^\dagger]^\dagger \in \mathbb{C}^{mL \times K}$ and $\boldsymbol{\zeta}_m = [\boldsymbol{\xi}_1^\dagger, \boldsymbol{\xi}_2^\dagger, \ldots, \boldsymbol{\xi}_m^\dagger]^\dagger \in \mathbb{C}^{mL}$. In addition, matrices $\overline{\mathbf{B}}_m^0 \in \mathbb{C}^{K \times mL}$ and $\overline{\boldsymbol{\Gamma}}_m^0 \in \mathbb{C}^{K \times mK}$ contain all the linear coefficients that multiply the received signals and the quantization noise vectors, respectively. Both of these matrices will be set iteratively below.

At APU 1, $\overline{\mathbf{B}}_1^0$ corresponds to the familiar MMSE matrix, while taking also into account the effect of the forthcoming compression. Thus, this matrix is given by

$$\overline{\mathbf{B}}_1^0 = \gamma_\epsilon \mathbf{Q} \hat{\mathbf{G}}_1^\dagger \left(\hat{\mathbf{G}}_1 \mathbf{Q} \hat{\mathbf{G}}_1^\dagger + \boldsymbol{\Omega}_1 \right)^{-1} \tag{20}$$

[2] In case where $m = M$, no further compression is needed.

[3] For simplicity purposes, in the present work, we assume that side information, which accompanies each soft estimation $\hat{\mathbf{s}}_m$, is transferred between APUs without distortion. However, we take into account the impact that side information has on each compression error variance via the number N of the total complex scalars transmitted, as seen in (17).

where $\gamma_\epsilon = (1+\epsilon)^{-1}$. Also, $\overline{\mathbf{\Gamma}}_1^0 = \mathbf{I}_K$, since the quantization noise is added after the combining process at APU 1.

Based on the above, at APU m, the incoming estimate $\bar{\mathbf{s}}_{m-1}$ is combined linearly with $\bar{\mathbf{y}}_m$, producing $\hat{\mathbf{s}}_m$ as

$$\hat{\mathbf{s}}_m = \mathbf{A}_m^0 \bar{\mathbf{s}}_{m-1} + \mathbf{B}_m^0 \bar{\mathbf{y}}_m \tag{21}$$

Consequently, upon the transmission of the above estimation via the mth finite-capacity link to the APU $m+1$, an additional quantization error vector $\mathbf{w}_m$ will be attached to it forming $\bar{\mathbf{s}}_m$, as seen in (16). By observing (16), (18) and (21), it derives that

$$\overline{\mathbf{B}}_m^0 = \left[\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0, \mathbf{B}_m^0\right] \tag{22}$$

$$\overline{\mathbf{\Gamma}}_m^0 = \left[\mathbf{A}_m^0 \overline{\mathbf{\Gamma}}_{m-1}^0, \mathbf{I}_K\right] \tag{23}$$

with $\mathbf{A}_m^0 \in \mathbb{C}^{K\times K}$, $\mathbf{B}_m^0 \in \mathbb{C}^{K\times L}$ being the combining matrices of each APU m. Then, optimality is achieved when

$$\mathbf{B}_m^0 = \gamma_\epsilon \left(\mathbf{Q} - \mathbf{\Lambda}_{m-1}\right) \hat{\mathbf{G}}_m^\dagger \mathbf{J}_m^{-1} \tag{24}$$

$$\mathbf{A}_m^0 = \gamma_\epsilon \left(\mathbf{I}_K - \mathbf{B}_m^0 \hat{\mathbf{G}}_m \gamma_\epsilon^{-1}\right) \mathbf{Q} \hat{\mathbf{H}}_{m-1}^\dagger \overline{\mathbf{B}}_{m-1}^{0\dagger} \mathbf{F}_{m-1}^{0-1} \tag{25}$$

where the factor γ_ϵ stems from optimizing $\bar{\mathbf{s}}_m$ rather than $\hat{\mathbf{s}}_m$. Also, matrices $\mathbf{F}_m^0$, $\mathbf{J}_m$ and $\mathbf{\Lambda}_m$ are defined as

$$\mathbf{F}_m^0 = \overline{\mathbf{B}}_m^0 \left(\hat{\mathbf{H}}_m \mathbf{Q} \hat{\mathbf{H}}_m^\dagger + \mathcal{K}_m\right) \overline{\mathbf{B}}_m^{0\dagger} + \overline{\mathbf{\Gamma}}_m^0 \mathcal{S}_m \overline{\mathbf{\Gamma}}_m^{0\dagger} \tag{26}$$

$$\mathbf{J}_m = \hat{\mathbf{G}}_m (\mathbf{Q} - \mathbf{\Lambda}_{m-1}) \hat{\mathbf{G}}_m^\dagger + \mathbf{\Omega}_m \tag{27}$$

$$\mathbf{\Lambda}_m = \mathbf{Q} \hat{\mathbf{H}}_m^\dagger \overline{\mathbf{B}}_m^{0\dagger} \mathbf{F}_m^{0-1} \overline{\mathbf{B}}_m^0 \hat{\mathbf{H}}_m \mathbf{Q} \tag{28}$$

with $\mathcal{K}_m = \text{diag}(\mathbf{\Omega}_1, \mathbf{\Omega}_2, \ldots, \mathbf{\Omega}_m)$ and $\mathcal{S}_m = \text{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_m)$. Also, notice that $\mathbf{\Sigma}_m$ represents the covariance matrix of each $\mathbf{w}_m$ and is determined as $\text{diag}(\mathbb{E}\{|w_{m,1}|^2\}, \mathbb{E}\{|w_{m,2}|^2\}, \ldots, \mathbb{E}\{|w_{m,K}|^2\})$. Finally, by observing (16), (18), (22), (23) and (26) it derives that $\mathbf{F}_m^0 = \mathbf{S}_m + \mathbf{\Sigma}_m$, where $\mathbf{S}_m = \mathbb{E}\{\hat{\mathbf{s}}_m \hat{\mathbf{s}}_m^\dagger\}$.

Proof: The proof that matrices $\mathbf{B}_m^0$ and $\mathbf{A}_m^0$ are the MMSE solution of (18) is given in Appendix A. ∎

In order for APU m to evaluate $\mathbf{B}_m^0$ and $\mathbf{A}_m^0$, apart from the local matrices $\hat{\mathbf{G}}_m$, $\mathbf{\Omega}_m$ and $\mathbf{\Sigma}_m$, it also requires previously created information. This information is acquired from the APU $m-1$ through the $K \times K$ Hermitian matrices $\overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q}$ and $\mathbf{F}_{m-1}^0$ and amounts to K^2 complex scalars. Thus, in each coherence interval τ_c, the overall information transmitted between two successive APUs, including the soft estimation vector $\hat{\mathbf{s}}_m$, totals to $N = K\tau_c$ complex scalars. Furthermore, these matrices can also be used to evaluate the $\text{SINR}_{m,k}^0$, for all $m \in [M]$ and $k \in [K]$, based on the channels and noises up to APU m as

$$\text{SINR}_{m,k}^0 = \frac{\left|\left[\overline{\mathbf{B}}_m^0 \hat{\mathbf{H}}_m \mathbf{Q}\right]_{k,k}\right|^2}{\rho_k \left[\mathbf{F}_m^0\right]_{k,k} - \left|\left[\overline{\mathbf{B}}_m^0 \hat{\mathbf{H}}_m \mathbf{Q}\right]_{k,k}\right|^2} \tag{29}$$

Notice that $\text{SINR}_{m,k}^0$ refers to the SINR of $\bar{s}_{m,k}$, considering also the mth quantization error. However, the final $\text{SINR}_{M,k}^0$ of each UE k at the CPU ($m = M$) expresses the quality of $\hat{s}_{M,k}$ and thus it is calculated as

$$\text{SINR}_{M,k}^0 = \frac{\left|\left[\overline{\mathbf{B}}_M^0 \hat{\mathbf{H}}_M \mathbf{Q}\right]_{k,k}\right|^2}{\rho_k \left[\mathbf{S}_M\right]_{k,k} - \left|\left[\overline{\mathbf{B}}_M^0 \hat{\mathbf{H}}_M \mathbf{Q}\right]_{k,k}\right|^2} \tag{30}$$

Proof: The proof that $\overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q}$ and $\mathbf{F}_{m-1}^0$ are Hermitian matrices is given in Appendix B, while in Appendix A we show how they can be evaluated at each APU. ∎

IV. The CnF Strategy

In the previous section, we proposed an uplink sequential processing algorithm that can be applied to radio stripe architectures and can optimally combine, at each APU m, the collected payload signal (compressed or not) with the compressed soft estimation vector $\bar{\mathbf{s}}_{m-1}$ that APU m has received from APU $m-1$. This processing method, in contrast with related works that regard conventional CF mMIMO layouts with imperfect FH [43], [44], [45], [46], [47], requires multiple compressions until all the signals reach the CPU. A key consequence of this successive signal acquisition and quantization procedure is that for some UEs $k \in \mathcal{P}_m$, where $\mathcal{P}_m \subseteq \{1, 2, \ldots, K\}$, the addition of new information to the scalar estimate $\bar{s}_{m-1,k}$ may not be able to counterbalance the quantization noise $w_{m,k}$ that will be added due to its compression, leading to

$$\text{SINR}_{m,k}^0 < \text{SINR}_{m-1,k}^0, \quad \forall k \in \mathcal{P}_m \tag{31}$$

In this case, it is advantageous, for each APU m, not to evaluate the renewed signal estimate $\hat{s}_{m,k}$ for these specific UEs $k \in \mathcal{P}_m$, but instead to relay their incoming ones $\bar{s}_{m-1,k}$ without adding new information or compression noise. It is important to point out that this comparison only has meaning in limited-capacity-link schemes, since when no quantization noise is present, it is always advantageous to add new information at each APU.

Remark 2: In practice, avoidance of the transmission chain has the advantage that no further compression noise will be attached to the already amplified signal, an idea that was first suggested in [1] for the N-LMMSE algorithm. In particular, by matching each UE k to a specific dynamic cooperation cluster $\mathcal{D}_k$ of APUs, based on the APS scheme of (31), led to the formation of the *user-centric* network in Fig. 2. This action, as will be shown later on, not only will ensure a higher throughput to most users, but also will greatly reduce the total latency of the system.

Next, we will analyze how this methodology can be applied to the OSPA algorithm. Since every APU m has access to both $\left\{\overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q}, \mathbf{F}_{m-1}^0\right\}$ and $\left\{\overline{\mathbf{B}}_m^0 \hat{\mathbf{H}}_m \mathbf{Q}, \mathbf{F}_m^0\right\}$ information, it can thus evaluate both $\text{SINR}_{m-1,k}^0$ and $\text{SINR}_{m,k}^0$, for every $k \in [K]$, which correspond to signal estimates $\bar{s}_{m-1,k}$ and $\bar{s}_{m,k}$, respectively. Then, for each UE k individually, it can transmit to APU $m+1$ the soft estimation that leads to the higher SINR value, meaning that it can either form and transmit $\hat{s}_{m,k}$ or relay $\bar{s}_{m-1,k}$, along with their respective side information. This action will not only benefit UE k, but

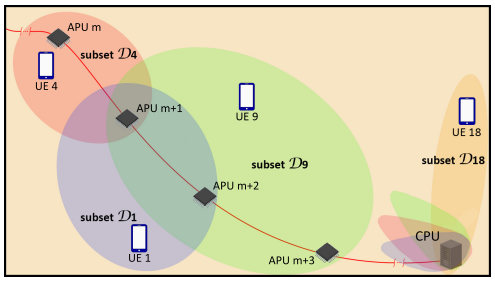


Fig. 2. Example of how different subsets $\mathcal{D}_k \subseteq \{1, 2, \ldots, M\}$ of APUs serve their assigned UEs k, where $k = 1, 4, 9, 18$, when the CnF strategy is applied.

also all the other users, as each soft estimation is a linear combination of all the other UEs' combined signals. To make that comparison, we define a diagonal matrix $\boldsymbol{\Theta}_m$ as follows

$$\boldsymbol{\Theta}_m = \text{diag}\left[\mathbf{1}\left\{\text{SINR}^0_{m,k} - \text{SINR}^0_{m-1,k}\right\}\right] \tag{32}$$

where $\mathbf{1}\{x\}$ is zero when $x < 0$ and one otherwise. Considering the above matrix, (18) can be rewritten as

$$\bar{\mathbf{s}}_m = (\mathbf{I}_K - \boldsymbol{\Theta}_m + \boldsymbol{\Theta}_m \mathbf{A}^0_m)\bar{\mathbf{s}}_{m-1} + \boldsymbol{\Theta}_m \mathbf{B}^0_m \bar{\mathbf{y}}_m + \boldsymbol{\Theta}_m \mathbf{w}_m \tag{33}$$

This means we may now redefine iterative operators $\overline{\mathbf{B}}^0_m$ and $\overline{\boldsymbol{\Gamma}}^0_m$ in the following way

$$\overline{\mathbf{B}}_m = \left[\mathbf{A}_m \overline{\mathbf{B}}_{m-1}, \mathbf{B}_m\right] \tag{34}$$

$$\overline{\boldsymbol{\Gamma}}_m = \left[\mathbf{A}_m \overline{\boldsymbol{\Gamma}}_{m-1}, \boldsymbol{\Theta}_m\right] \tag{35}$$

where matrices $\mathbf{A}^0_m$ and $\mathbf{B}^0_m$ are now given as

$$\mathbf{A}_m = \mathbf{I}_K - \boldsymbol{\Theta}_m + \boldsymbol{\Theta}_m \mathbf{A}^0_m \tag{36}$$

$$\mathbf{B}_m = \boldsymbol{\Theta}_m \mathbf{B}^0_m \tag{37}$$

Additionally, as seen in Appendix A, the construction of matrices $\mathbf{F}^0_m$ and $\overline{\mathbf{B}}^0_m \hat{\mathbf{H}}_m \mathbf{Q}$ also involves the matrices $\mathbf{B}^0_m$, $\mathbf{A}^0_m$ and $\boldsymbol{\Sigma}_m$. Hence, according to (34) and (35), side information matrices will also have to be recalculated using this time the redefined matrices $\mathbf{B}_m$, $\mathbf{A}_m$ and $\boldsymbol{\Theta}_m \boldsymbol{\Sigma}_m$. These new matrices, now denoted as $\overline{\mathbf{B}}_m \hat{\mathbf{H}}_m \mathbf{Q}$ and $\mathbf{F}_m$, are the ones that APU $m+1$ will eventually receive, from APU m, and use for its local computations. Finally, since APU 1 and CPU always serve all UEs, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_M$ are always identity matrices.

Following the above manipulations, the $\text{SINR}^0_{m,k}$ in (29) is replaced by[4]

$$\text{SINR}_{m,k} = \frac{\left|\left[\overline{\mathbf{B}}_m \hat{\mathbf{H}}_m \mathbf{Q}\right]_{k,k}\right|^2}{\rho_k \left[\mathbf{F}_m\right]_{k,k} - \left|\left[\overline{\mathbf{B}}_m \hat{\mathbf{H}}_m \mathbf{Q}\right]_{k,k}\right|^2} \tag{38}$$

The above procedure may be iterated all the way to the CPU where the decoding of the signal takes place. By treating all noise sources as being zero-mean complex Gaussian and independent of every message signal q_k [42], we can evaluate

[4]Notice that the $\text{SINR}_{m,k}$ of (38) is the actual SINR value that APU $m+1$ will compare to $\text{SINR}^0_{m+1,k}$, for all UE $k \in [K]$, in order to derive $\boldsymbol{\Theta}_{m+1}$.

Algorithm 1 The Application of CnF to OSPA

1) **Begin:** set $\hat{\mathbf{s}}_0 = \mathbf{0}$, $\overline{\mathbf{B}}^0_0 \hat{\mathbf{H}}_0 \mathbf{Q} = \mathbf{0}$, $\mathbf{F}^0_0 = \mathbf{I}_K$
2) **for** $m = 1 : M$ **do**
   a) Compute $\boldsymbol{\Lambda}_{m-1}$, $\mathbf{J}_m$, $\mathbf{B}^0_m$, $\mathbf{A}^0_m$, $\overline{\mathbf{B}}^0_m \hat{\mathbf{H}}_m \mathbf{Q}$, $\{\hat{\mathbf{s}}_m | \mathbf{B}^0_m, \mathbf{A}^0_m\}$, $\mathbf{F}^0_m$
   b) **if** $m = 1$ **then** transmit $\hat{\mathbf{s}}_1$, $\overline{\mathbf{B}}^0_1 \hat{\mathbf{H}}_1 \mathbf{Q}$, $\mathbf{F}^0_1$
   elseif $1 < m < M$ **then**
      i) Compute $\text{SINR}_{m-1,k}$ and $\text{SINR}^0_{m,k}$ for every $k \in [K]$
      ii) Compute $\boldsymbol{\Theta}_m$ using the results of step (i)
      iii) Using $\boldsymbol{\Theta}_m$, compute $\mathbf{B}_m$, $\mathbf{A}_m$, $\mathbf{F}_m$, $\overline{\mathbf{B}}_m \hat{\mathbf{H}}_m \mathbf{Q}$, $\{\hat{\mathbf{s}}_m | \mathbf{B}_m, \mathbf{A}_m\}$
      iv) Transmit $\{\hat{\mathbf{s}}_m | \mathbf{B}_m, \mathbf{A}_m\}$, $\overline{\mathbf{B}}_m \hat{\mathbf{H}}_m \mathbf{Q}$, $\mathbf{F}_m$
   end if
   end for
3) **Output:** $\hat{\mathbf{s}}_M$, $\overline{\mathbf{B}}^0_M \hat{\mathbf{H}}_M \mathbf{Q}$ and $\mathbf{S}_M$

the achievable per-user SE_k (lower bound of the ergodic channel capacity of UE k) as [5]

$$\text{SE}_k = \frac{\tau_d}{\tau_c} \mathbb{E}\left\{\log_2(1 + \text{SINR}_{M,k})\right\} \tag{39}$$

where the expectation is over the channel estimations and the decision, at each APU m, to either transmit $\hat{s}_{m,k}$ or relay $\bar{s}_{m-1,k}$. Also, since every UE is always served from the Mth APU, the $\text{SINR}_{M,k}$ is calculated exactly as in (30). The above strategy is summarized in the form of a pseudocode in Algorithm 1.

V. The Error Minimization Strategy

In Section III, we assumed that the compression of each soft estimation element $\hat{s}_{m,k}$, for $k = 1, 2, \ldots, K$, was performed independently, thus neglecting the correlation between them. This allowed us to use a simple expression for the quantization error covariance matrix $\boldsymbol{\Sigma}_m$, which in turn was used in the derivation of the optimal combining matrices $\mathbf{B}^0_m$ and $\mathbf{A}^0_m$, as seen in Appendix A. Moreover, after applying the CnF strategy, the final covariance matrix of the soft estimation $\hat{\mathbf{s}}_m$ that APU m would ultimately transmit to APU $m+1$ could differ from the one that was used to derive $\boldsymbol{\Sigma}_m$. Nevertheless, due to the simplified structure of the aforementioned quantizer, that change did not affect the diagonal elements of $\boldsymbol{\Sigma}_m$ and thus we did not have to calculate them over again after the application of CnF.

Since the MSE_m expression for the mth APU in (50) depends on $\boldsymbol{\Sigma}_m$ through the $\text{tr}(\boldsymbol{\Sigma}_m)$ and on the previous APUs' error covariance matrices through the $\text{tr}(\mathbf{A}^0_m \overline{\boldsymbol{\Gamma}}^0_{m-1} \mathcal{S}_{m-1} \overline{\boldsymbol{\Gamma}}^{0\dagger}_{m-1} \mathbf{A}^{0\dagger}_m)$, it is possible to further reduce it as well as to reduce every upcoming $\text{MSE}_{m'}$, where $m' > m$, by minimizing the term $\text{tr}(\boldsymbol{\Sigma}_m)$, subject of course to the limited-capacity constraint of the corresponding mth link. The procedure for this minimization appears in the next Proposition.

Proposition 1: Let $r_{m1} \geq r_{m2} \geq \cdots \geq r_{mK}$ be the eigenvalues and $\mathbf{U}_m$ the unitary matrix with the corresponding

eigenvectors of $\mathbf{S}_m$, where $\mathbf{S}_m$ can also be expressed as $\mathbf{S}_m = \mathbf{U}_m \text{diag}(r_{m1}, r_{m2}, \ldots, r_{mK}) \mathbf{U}_m^{\dagger}$. Then, it can be shown that the minimum $\text{tr}(\mathbf{\Sigma}_m)$, under the constraint $\mathcal{R} = \frac{1}{\tau_c} \log_2 \det(\mathbf{S}_m \mathbf{\Sigma}_m^{-1} + \mathbf{I}_K)$, is achieved when the matrices $\mathbf{\Sigma}_m$ and $\mathbf{S}_m$ are simultaneously diagonalizable in the common eigenvector basis (i.e. $\mathbf{\Sigma}_m = \mathbf{U}_m \mathbf{\Delta}_m \mathbf{U}_m^{\dagger}$). In this case, the minimum $\text{tr}(\mathbf{\Sigma}_m)$ reduces to $\sum_{k=1}^{K} x_{mk}$, where $x_{m1} \geq x_{m2} \geq \cdots \geq x_{mK}$ are the diagonal elements of $\mathbf{\Delta}_m$ that are computed as

$$x_{mk} = \frac{-r_{mk} + \sqrt{r_{mk}^2 + 4 r_{mk} \lambda}}{2}, \quad \forall k \in [K] \tag{40}$$

with the coefficient λ to be obtained as

$$\mathcal{R} = \frac{1}{\tau_c} \sum_{i=1}^{K} \log_2 \left(1 + \frac{2}{\sqrt{1 + \frac{4\lambda}{r_{mi}}} - 1} \right) \tag{41}$$

Proof: As $\text{tr}(\mathbf{\Sigma}_m)$ has K degrees of freedom, the minimization of that function leads to various covariance matrices $\mathbf{\Sigma}_m$. To distinguish the appropriate solution for our case, we will exploit the limited-capacity restriction that each link between two successive APUs is subjected to. Namely, the minimum rate $\mathcal{R}$ needed for the transmission of each vector $\hat{\mathbf{s}}_m$ from APU m to APU $m+1$ is associated with any compression variance $\mathbf{\Sigma}_m$ as

$$\begin{aligned} \mathcal{R} &= \frac{1}{\tau_c} \log_2 \det \left(\mathbf{S}_m \mathbf{\Sigma}_m^{-1} + \mathbf{I}_K \right) \\ &= \frac{1}{\tau_c} \log_2 \det \left(\mathbf{\Sigma}_m^{-1} + \mathbf{S}_m^{-1} \right) + \frac{1}{\tau_c} \log_2 \det \left(\mathbf{S}_m \right) \end{aligned} \tag{42}$$

The above expression acts as a constraint for the aforementioned degrees of freedom, allowing us to conclude to a unique solution. Using the method of Lagrange multipliers, that problem can be formulated as follows

$$\begin{aligned} &\min_{\{x_{m1}, x_{m2}, \ldots, x_{mK}, \lambda\}} \quad \mathcal{L}(x_{mk}, \lambda) \\ &= \sum_{j=1}^{K} x_{mj} + \lambda \left[\ln \det \left(\mathbf{S}_m \mathbf{\Sigma}_m^{-1} + \mathbf{I}_K \right) - \tau_c \mathcal{R} \ln 2 \right] \end{aligned} \tag{43}$$

where $\lambda \geq 0$ is the Lagrange multiplier for the explicit constraint of (42). Moreover, since $\mathbf{\Sigma}_m$ and $\mathbf{S}_m$ are positive semi-definite matrices, it is obvious that $x_{mk} + r_{mk} \geq 0$, for every $k \in [K]$. Hence, term $\ln \det(\mathbf{S}_m \mathbf{\Sigma}_m^{-1} + \mathbf{I}_k)$ is lower bounded [52] as

$$\sum_{i=1}^{K} \ln \left(\frac{r_{mi}}{x_{mi}} + 1 \right) \leq \ln \det(\mathbf{S}_m \mathbf{\Sigma}_m^{-1} + \mathbf{I}_K) \tag{44}$$

Comparing (43) and (44), it derives that the minimum $\mathcal{L}(x_{mk}, \lambda)$ is achieved in case where $\mathbf{\Sigma}_m$ and $\mathbf{S}_m$ are simultaneously diagonalizable. In that occasion, the minimization problem of (43) reduces to

$$\begin{aligned} &\min_{\{x_{m1}, x_{m2}, \ldots, x_{mK}, \lambda\}} \quad \mathcal{L}(x_{mk}, \lambda) \\ &= \sum_{j=1}^{K} x_{mj} + \lambda \left[\sum_{i=1}^{K} \ln \left(\frac{r_{mi}}{x_{mi}} + 1 \right) - \tau_c \mathcal{R} \ln 2 \right] \end{aligned} \tag{45}$$

To further minimize the Lagrangian function, we will find the stationary points of (45) by differentiation. That is done by setting all partial derivatives equal to zero, leading to the following $K+1$ equation system

$$\begin{cases} x_{mk}^2 + r_{mk} x_{mk} - r_{mk} \lambda = 0, \quad \forall k \in [K] \\ \sum_{i=1}^{K} \log_2 \left(\frac{r_{mi}}{x_{mi}} + 1 \right) - \tau_c \mathcal{R} = 0 \end{cases} \tag{46}$$

By solving the above equation system for every non-negative eigenvalue x_{mk}, where $k \in [K]$, yields (40) and (41). ■

Notice that since $\mathcal{R}$ is the transmission rate of the K complex symbols that represent each $\hat{\mathbf{s}}_m$, the total capacity (in bits/s/Hz) that is allocated for the transmission of all these soft estimations along with their respective side information, in each coherence interval τ_c, is calculated as $C_{apu} = \mathcal{R} N K^{-1}$. The aforementioned minimization technique is summarized in the form of a pseudocode in Algorithm 2.

Algorithm 2 Compression Error Covariance Trace Minimization

1) **Begin:** set $\hat{\mathbf{s}}_0 = \mathbf{0}$, $\overline{\mathbf{B}}_0^0 \hat{\mathbf{H}}_0 \mathbf{Q} = \mathbf{0}$, $\mathbf{F}_0^0 = \mathbf{I}_K$, $\gamma_\epsilon = 1$
2) **for** $m = 1 : M$ **do**
   a) Compute $\mathbf{\Lambda}_{m-1}$, $\mathbf{J}_m$, $\mathbf{B}_m^0$, $\mathbf{A}_m^0$, $\overline{\mathbf{B}}_m^0 \hat{\mathbf{H}}_m \mathbf{Q}$, $\{\hat{\mathbf{s}}_m | \mathbf{B}_m^0, \mathbf{A}_m^0\}$ and $\mathbf{S}_m$
   b) Using $\mathbf{S}_m$, compute the optimal $\mathbf{\Sigma}_m$
   c) Compute $\mathbf{F}_m^0 = \mathbf{S}_m + \mathbf{\Sigma}_m$
   d) **if** $m < M$ **then** transmit $\hat{\mathbf{s}}_m$, $\overline{\mathbf{B}}_m^0 \hat{\mathbf{H}}_m \mathbf{Q}$, $\mathbf{F}_m^0$

   end for
3) **Output:** $\hat{\mathbf{s}}_M$, $\overline{\mathbf{B}}_M^0 \hat{\mathbf{H}}_M \mathbf{Q}$ and $\mathbf{S}_M$

Remark 3: It is worth pausing for a moment and comparing the optimization procedures discussed so far. Starting from Section III, we minimized the total MSE_m in (50) with respect to the matrices $\mathbf{A}_m^0$ and $\mathbf{B}_m^0$, including the total square error due to compression, i.e. $\text{tr}(\mathbf{\Sigma}_m)$, which also depends on these matrices. Next, in Section IV, the resulting SINRs for each user at the mth APU are compared to the respective SINRs achieved at APU $m-1$, while afterwards the data corresponding to the highest per-user SINR is sent to the next APU. In the present section, to minimize the total MSE $\mathbb{E}\{\|\bar{\mathbf{s}}_m - \mathbf{q}\|^2\}$ appearing in (50), we proceed in two steps. First, we minimize the MSE $\mathbb{E}\{\|\hat{\mathbf{s}}_m - \mathbf{q}\|^2\}$ with respect to $\mathbf{A}_m^0$ and $\mathbf{B}_m^0$, which essentially is the MSE_m in (50) in the absence of the quantization noise term $\text{tr}(\mathbf{\Sigma}_m)$. Then, we minimize that noise term over all users jointly, subject to the capacity constraint of the link between APUs m and $m+1$, i.e. $\mathcal{R} = \frac{1}{\tau_c} \log_2 \det(\mathbf{S}_m \mathbf{\Sigma}_m^{-1} + \mathbf{I}_K)$, where $\mathbf{S}_m$ is the covariance matrix of $\hat{\mathbf{s}}_m$. However, in this case, we do not apply the CnF strategy, since the optimization over the $\mathbf{\Sigma}_m$ depends on the statistics of all UEs' signals. Nevertheless, we shall see that this method performs better than no optimization at all.

VI. Performance Analysis

A. Complexity and Latency Issues

In this subsection, we discuss the complexity of the applied algorithms and their service latency. Both of these aspects

will be analyzed within the duration of a single coherence interval τ_c.

Starting with the main algorithm of Section III, its complexity can be analyzed as follows. At first, the $\mathbf{F}_{m-1}^0 \in \mathbb{C}^{K\times K}$ matrix is inverted and then the $\mathbf{\Lambda}_{m-1} \in \mathbb{C}^{K\times K}$ matrix is formed, with $O(K^3)$ complexity. This is also the complexity for the formulation of the matrices $\mathbf{A}_m^0$ and $\mathbf{B}_m^0$, as $K \gg L$. Subsequently, notice that for each temporal instance, the final K-dimensional data vector estimate $\hat{\mathbf{s}}_m$, at the mth APU, is the result of multiplication of the fixed matrices $\mathbf{A}_m^0$ and $\mathbf{B}_m^0$ on the vectors $\bar{\mathbf{s}}_{m-1}$ and $\bar{\mathbf{y}}_m$ respectively, as seen in (21). Hence, since $K \gg L$, the complexity of these operations for the entire payload period τ_d is $O(K^2\tau_d)$, a result that is typical for similar studies [21]. Hence, the total complexity ends up to be $O(K^2\tau_d + K^3)$. However, notice that since $\tau_d > K$ for all instances we are studying, the complexity for the formation of the matrices $\mathbf{A}_m^0$ and $\mathbf{B}_m^0$ is subdominant. Additionally, when CnF (Algorithm 1) is applied to the OSPA, the complexity remains the same, as the $K \times K$ $\mathbf{\Theta}_m$ matrix that multiplies $\mathbf{A}_m^0$ and $\mathbf{B}_m^0$, as seen in (36) and (37), is diagonal.

Moving on to Algorithm 2, it should be mentioned that has a similar complexity with that of Algorithm 1, with one additional evaluation, namely the optimization of the $K \times K$ covariance matrix $\mathbf{\Sigma}_m$ of the quantization noise $\mathbf{w}_m$. This optimization involves the diagonalization of $\mathbf{S}_m$, which has $O(K^3)$ complexity, and subsequently the optimization of its eigenvalues, using (40) and (41), with $O(K)$ complexity.

An alternative sequential processing concept that can reduce the complexity at each APU is this of the MRC algorithm. In that case, the estimate $\hat{\mathbf{s}}_m^{\text{mrc}}$ is given as

$$\hat{\mathbf{s}}_m^{\text{mrc}} = \bar{\mathbf{s}}_{m-1}^{\text{mrc}} + \hat{\mathbf{G}}_m^{\dagger}\mathbf{\Omega}_m^{-1}\bar{\mathbf{y}}_m = \hat{\mathbf{H}}_m^{\dagger}\mathcal{K}_m^{-1}\bar{\boldsymbol{\upsilon}}_m + \sum_{i=1}^{m-1}\mathbf{w}_i \quad (47)$$

It is easy to see that the above corresponds to the updating algorithm appearing in (21), where $\mathbf{A}_m^0 = \mathbf{I}_K$, $\mathbf{B}_m^0 = \hat{\mathbf{G}}_m^{\dagger}\mathbf{\Omega}_m^{-1}$ and hence is by construction suboptimal to the MMSE solution discussed in Section III. However, since APU m does not require any previous information in order to form the local combining matrix, the FH signaling in each link between the APUs reduces from $N = K\tau_c$ to $N = K\tau_d$ complex symbols. Additionally, since $\mathbf{\Omega}_m$ is an $L\times L$ diagonal matrix and $\hat{\mathbf{G}}_m^{\dagger}$ is a $K \times L$ matrix, the complexity here is $O(KL\tau_d)$, which is substantially less than this of the OSPA, as $K \gg L$ in most cases.

Next, we discuss about the issue of latency, which is the total time T_{tot} that is required in order for the entire signal of each UE, transmitted within a coherence interval τ_c, to reach the CPU so that it can be fully decoded. Assuming that the processing time at each APU is t_p (also considering the combining delay at the CPU) and that the coherence time is T_c, then the total delay time of each UE will be $T_{tot} = T_c + Mt_p$.

The above hold for both the standard OSPA algorithm, discussed in Section III, as well as the Algorithm 2, which was introduced in Section V. However, in the case of CnF (Algorithm 1), the total delay time of each UE differentiates, as it is jointly dependent on their distance from the CPU and the given radio stripe's capacity (commentary of figures 4 and 8). More specifically, when the SINR of a UE k is maximized, then their soft estimation can be directly transmitted to the CPU, thus reducing their service latency to $T_{tot}^k = T_c + (m_k + 1)t_p$, where $m_k + 1 \leq M$. Notice that APU m_k is the final APU (except the CPU that always adds information) that adds information to the soft signal estimate of UE k, i.e. is the APU that marks the transmission point of that specific UE, as demonstrated in Fig. 4. Hence, the UEs located far from the CPU and whose performance is maximized from the very first APUs, gain lower latency than the UEs close to it. This behaviour can be seen clearly in Fig. 8, where the number of users served by each APU falls linearly with distance.

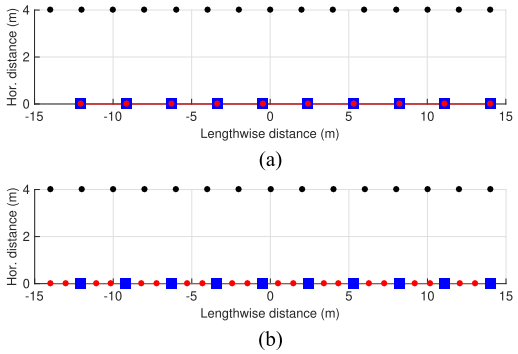


Fig. 3. Radio stripe setups used for the simulation, where scheme (a) realizes Fig. 1a and scheme (b) realizes Fig. 1b. Black dots represent the UEs, red dots the APs and blue squares the APUs.

B. Numerical Results

We consider the two geometries of Fig. 1, both placed at a fixed horizontal distance of 4 m away from the UEs, as shown in Fig. 3. The users have an intermediate distance of 2 m between each other, while the radio stripe's components are equally distributed lengthwise of them, so that they occupy the exact same space extent.

To evaluate the performance of all the above mentioned algorithms, under the limited-capacity restriction, we use the 3GPP Urban Microcell model [53, Table B.1.2.1-1], which matches well with outdoor radio stripe applications and which calculates the large-scale fading between the kth UE and the lth AP of the mth APU as

$$\beta_{mkl}(\text{dB}) = -36.7\log_{10}(d_{mkl}) - 26\log_{10}(f_c) - 22.7 \quad (48)$$

where f_c (GHz) is the carrier frequency and d_{mkl} (m) is the direct distance between UE k and the lth AP of the mth APU. Moreover, noise power σ_n^2 is given as

$$\sigma_n^2\ (\text{W}) = \text{bandwidth} \times k_B \times T_0 \times \text{noise figure} \quad (49)$$

where $k_B = 1.384\times10^{-23}$ (Joule per Kelvin) is the Boltzmann constant and $T_0 = 290$ (K) the noise temperature. The rest of the parameters are summarized in Table I.

In Fig. 4, we numerically verify the uplink SE of (39) that is achieved for two specific UEs, located roughly at one-third and at two-thirds of the total layout distance, at each APU. The curves have iteratively been derived using 500 independent random-fading channel instantiations. Specifically, we plot the

TABLE I
SUMMARY OF PARAMETERS USED IN THE SIMULATION

Parameter	Value
Carrier frequency f_c	2 GHz
Bandwidth B	20 MHz
Noise figure	9 dB
Transmit power ρ_k	100 mW
Coherence block τ_c	180 samples
Coherence bandwidth B_c	100 kHz
Coherence time T_c	1.8 ms

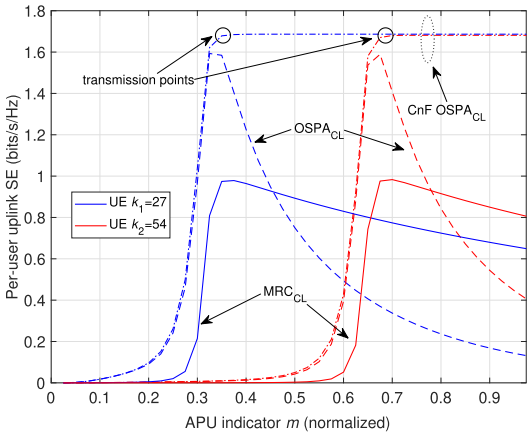
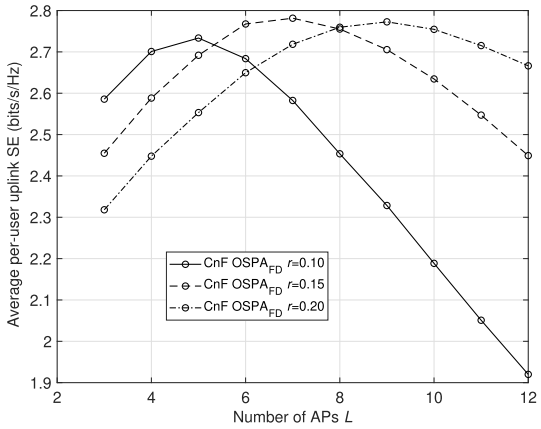
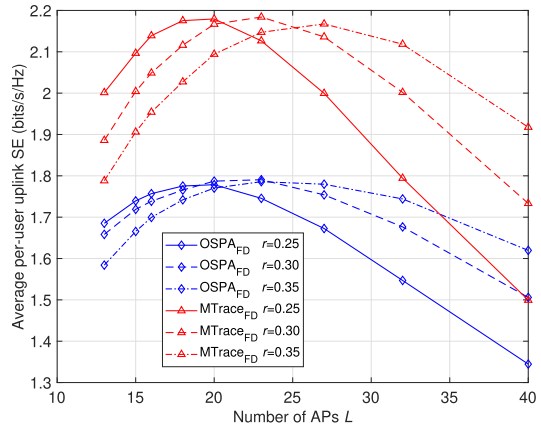


Fig. 4. The uplink SE achieved for UEs $k_1 = 27$ and $k_2 = 54$ at each APU m (normalized value), when the MRC, OSPA and CnF OSPA algorithms are applied to the setup of Fig. 3a. Here, $K = 80$, $M = 40$, $L = 4$ and $C = 300$ bits/s/Hz.



(a) Utilization of CnF OSPA_{FD} algorithm.



(b) Utilization of OSPA_{FD} and $\text{MTrace}_{\text{FD}}$ algorithms.

Fig. 5. Average per-user uplink throughput achieved as L varies, when the CnF OSPA, OSPA and MTrace algorithms are applied to the setup of Fig. 3b, for three different capacity allocation ratios r. Here, $K = 80$, $C = 500$ bits/s/Hz and product $M \times L \approx 160$ (fixed).

SE for the plain OSPA, CnF OSPA and MRC algorithms for a radio stripe setup with $M = 40$ APUs, each one having $L = 4$ collocated APs. Initially, the achievable throughput of each UE increases as their soft signal estimate acquire additional and higher quality signal. However, in contrast to infinite-capacity scenarios [21] where this rising trajectory continues until the CPU for all UEs, at some APU (specific for each UE), the performance of the plain OSPA and MRC algorithms, starts to deteriorate due to increased compression noise. In fact, compression noise, for the case of the plain OSPA, is so severe that even the suboptimal MRC algorithm manages to surpass its SE, as it requires much less FH signaling in order to function. Nonetheless, by avoiding unnecessary compressions, CnF OSPA can sustain the peak performance of each UE, thus preserving the benefits of the OSPA and rendering MRC inferior in any case. An additional key observation of this plot is that the CnF strategy substantially reduces the latency of some UEs, especially of those who are far from the CPU. This is achieved by transmitting to the CPU the current soft estimate of each UE when it reaches its peak performance (transmission point). These transmitted "optimum" estimates are also used from all the intermediate APUs in order to assist in the formation of all the remaining soft estimations.

Fig. 5 illustrates the average per-user uplink SE that is achieved as the number of APs varies (the total length of the radio stripe is kept fixed), when the CnF OSPA, OSPA and MTrace algorithms are applied to the fully distributed (FD) setup of Fig.3b, for 3 different capacity allocation ratio r scenarios. The number of UEs is set to $K = 80$, the capacity to $C = 10$ Gbps [e.g. optical multi-mode 3 (OM3) fiber] and the product $M \times L \approx 160$ is kept fixed. The results demonstrate that, as the number of APs increases, every curve reaches a maximum point beyond which any AP addition leads to a performance drop. That occurs because on one hand a scarce number of APs begets poor interference mitigation, but on the other hand a relatively large number confines the available transmission rate of both pilot and payload signals, thus inducing stronger compression errors p_m and d_m [see eq. (7)]. As curves in Fig. 5a and 5b indicate, the optimum trade-off between these two counterbalancing effects is obtained when $r = 0.15$ and $L = 7$ ($M = 23$) for the case of the CnF OSPA_{FD} and when $r = 0.30$ and $L = 23$ ($M = 7$) for the cases of the OSPA_{FD} and $\text{MTrace}_{\text{FD}}$.

In Fig. 6, we consider the same setups as in Fig. 5, but this time fixed to $M = 7$ APUs and $L = 23$ APs for the cases of the OSPA_{FD} and $\text{MTrace}_{\text{FD}}$ algorithms and $M = 23$ APUs and $L = 7$ APs for the case of the CnF OSPA_{FD} (optimal settings of Fig. 5). The plot shows the impact of various capacity allocation ratios on the average per-user uplink SE, when the radio stripe capacity is set to $C = 10$ Gbps. The curves validate the results of Fig. 5, as the maximum average

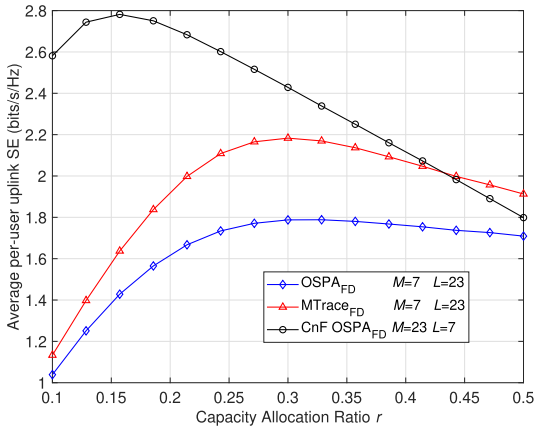


Fig. 6. Average per-user uplink throughput achieved as capacity allocation ratio r varies, when the CnF OSPA ($M = 23$, $L = 7$), OSPA ($M = 7$, $L = 23$) and MTrace ($M = 7$, $L = 23$) algorithms are applied to the setup of Fig. 3b. Here, $K = 80$ and $C = 500$ bits/s/Hz.

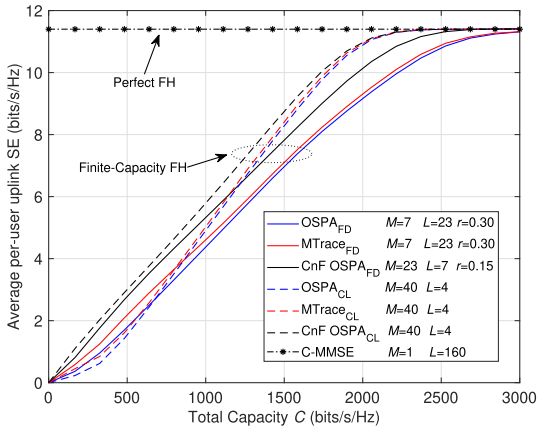


Fig. 7. Average per-user uplink throughput achieved as total capacity C varies, when the CnF OSPA, OSPA and MTrace algorithms are applied to both setups of Fig. 3. For the cases of the distributed-AP radio stripes, values of M, L and r have been chosen optimally based on the aforementioned results. Here, $K = 80$ and product $M \times L \approx 160$ (fixed).

performance is achieved when r, for each case separately, lies quite close to the previously found optimum values.

In Fig. 7, we observe the effect of the FH capacity C on the average per-user uplink SE, for six different radio stripe arrangements (all have product $M \times L \approx 160$ fixed for equality reasons) that serve $K = 80$ UEs. The solid curves refer to the optimized and fully distributed setups of Fig. 6, each one corresponding to one of the three algorithms. The dashed curves refer to three collocated-AP (CL) radio stripe layouts (see Fig. 3a) that share the same physical properties, yet each one utilizing a different algorithm. Furthermore, as this architecture does not allow for a big number of collocated APs, L is always set to 4. The plot stresses the great potential of the OSPA algorithm, as for a relatively large capacity (when quantization errors are small and thus capacity is considered to be unbounded), it behaves equivalently to the centralized MMSE (C-MMSE) arrangement, provided that both setups have the same AP distribution. However, the most important highlight of the plot is that the CnF OSPA algorithm, combined with the collocated-AP setup (black dashed curve), can achieve superior performance than any other scheme, irrespective of the available capacity of the radio stripe. This derives from two main reasons: The first is the aggressive dynamic cooperation clustering that the CnF strategy enables, which can effectively reduce the amount of redundant compressions throughout the sequential procedure. In addition to that, classic radio stripe topology of Fig. 3a ensures the elimination of p_m and d_m compression errors, a robust combination that leads to the aforementioned result. Nonetheless, distributed-AP radio stripes can still be useful in low-capacity regimes, if the MTrace and plain OSPA are the only options.

In Fig. 8, we investigate the percentage of UEs that each APU m serves, for the 3 collocated-AP radio stripe cases (dashed curves) of Fig. 7 and for 3 different capacity scenarios. The number of UEs considered for the results is also set to $K = 80$. It is evident that the less available capacity, the more drastic CnF becomes and the more rapidly starts to diminish the number of UEs that each APU serves. In practice, that means that when a UE k' is served by the APU that offers them the highest value of SINR (i.e. the APU that is

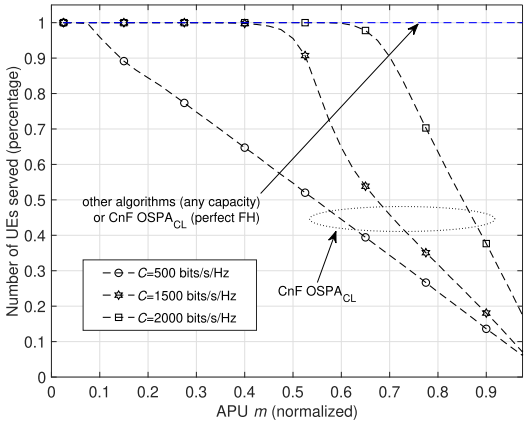


Fig. 8. Percentage of users that each APU serves as a function of each APU m (normalized value), when the CnF OSPA, OSPA and MTrace algorithms are applied to the setup of Fig. 3a, for three different capacity scenarios. Here, $K = 80$, $M = 40$ and $L = 4$.

closer to them), then at some point, all the subsequent APUs (except the CPU) will always offer them a worse trade-off between payload information and compression noise. Thus, from this point on (specific APU for each user), the CnF strategy prevents all the subsequent APUs from providing service to UE k' (as well as to all the other UEs prior to UE k'), a fact that justifies the strictly decreasing behaviour of every curve. This is exactly the point, also mentioned as transmission point in Fig. 4, that the soft estimate of UE k' has reached its maximum potential and thus can be immediately transmitted to the CPU, improving this way the latency of the users, especially of those who are far way from the CPU. Besides UE and APU positioning, that critical point strongly depends on the contextual capacity of the radio stripe, since it determines how severe the impact of the quantization noise on all forming SINRs will be. Nevertheless, due to the small-scale fading randomness, that linear behaviour may not rigorously apply for each individual scenario.

Finally, Fig. 9 demonstrates the average per-user uplink SE achieved as the number of UEs and APUs grows simultaneously, when all the three algorithms are applied to the

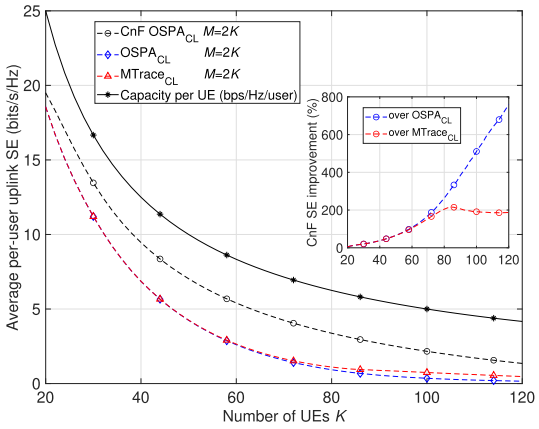


Fig. 9. Average per-user uplink throughput achieved as K and M vary, when the CnF OSPA, OSPA and MTrace algorithms are applied to the setup of Fig. 3a. Small plot demonstrates the percentage performance improvement that the CnF OSPA$_{\text{CL}}$ gains over the other two algorithms. Here, $L = 4$ and $C = 500$ bits/s/Hz.

arrangement of Fig. 3a. Here, the capacity is set to $C = 10$ Gbps and the number of APs to $L = 4$. As one would expect, average performance drops as the entire system scales up, an effect that is provoked by both the surging UE data (due to K increase) and mandatory compressions (due to M increase). The key highlight in this plot is the usefulness of the CnF OSPA$_{\text{CL}}$, which by effectively reducing the number of redundant compressions, manages to offer better results than the other two algorithms, particularly when the system is quite large. For instance, in the extreme case where $M = 200$ and $K = 100$, CnF ensures 6.1 times higher throughput than the plain OSPA$_{\text{CL}}$ and 2.9 times higher throughput than the MTrace$_{\text{CL}}$. Hence, this enhancing ability of the CnF strategy pushes the uplink performance of the average user closer, as compared to the other algorithms, to the optimal solid curve (ideal no-noise-and-error throughput) regardless of the layout's parameters.

VII. Conclusion

This paper models the impact of finite-capacity on the uplink performance of a cell-free massive MIMO topology that uses a consecutively-implemented fronthaul network with radio stripes. Within this context, we have analyzed a sequential processing algorithm that can optimally suppress interference-plus-noise, including compression noise, locally at each APU. To further mitigate the effect of quantization distortion, we have proposed two novel and capacity-efficient strategies, which can be used jointly with the OSPA and which are capable of augmenting its performance under any scenario tested. Moreover and in parallel with the classic radio stripe scheme, we have also examined an alternative arrangement, which, by leveraging its distributed-AP structure, could potentially lead to more optimized results.

Numerical simulations conclude that when the OSPA is combined with the CnF strategy, it has the potential to outperform any other algorithm tested, offering improved throughput and reduced latency to the majority of the network users, especially to those who lie far from the CPU. That is achieved thanks to the *user-centric* DCC framework that CnF enables, which by choosing the most suitable APUs to serve each UE, avoids redundant compressions, an action that it is rather vital in extensive layouts or, in general, when the radio stripe's capacity is scarce when compared with the contextual FH requirements. Finally, our results showcase the importance of combining the CnF OSPA with the classic radio stripe setup, namely the one that includes collocated APs, regardless of the under study implementation.

Appendix A
Optimal Combining Matrices Derivation

In this section we will prove that the matrices $\mathbf{B}_m^0$ and $\mathbf{A}_m^0$ are optimal in the sense of minimizing the mean-squared error between the compressed soft estimation $\bar{\mathbf{s}}_m$ and the message vector $\mathbf{q}$. Using (18), the $\text{MSE}_m = \mathbb{E}\left\{\|\bar{\mathbf{s}}_m - \mathbf{q}\|^2\right\}$ can be analyzed as follows

$$\begin{aligned}
\text{MSE}_m &= \mathbb{E}\left\{\|\overline{\mathbf{B}}_m^0 \bar{\boldsymbol{\upsilon}}_m + \overline{\boldsymbol{\Gamma}}_m^0 \boldsymbol{\omega}_m - \mathbf{q}\|^2\right\} \\
&= \mathbb{E}\Big\{\|\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \bar{\boldsymbol{\upsilon}}_{m-1} + \mathbf{B}_m^0 \bar{\mathbf{y}}_m \\
&\qquad + \mathbf{A}_m^0 \overline{\boldsymbol{\Gamma}}_{m-1}^0 \boldsymbol{\omega}_{m-1} + \mathbf{w}_m - \mathbf{q}\|^2\Big\} \\
&= \text{tr}\left(\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q} \hat{\mathbf{H}}_{m-1}^\dagger \overline{\mathbf{B}}_{m-1}^{0\dagger} \mathbf{A}_m^{0\dagger}\right) \\
&\quad + \text{tr}\left(\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \mathcal{K}_{m-1} \overline{\mathbf{B}}_{m-1}^{0\dagger} \mathbf{A}_m^{0\dagger}\right) \\
&\quad + \text{tr}\left[\mathbf{B}_m^0 \left(\hat{\mathbf{G}}_m \mathbf{Q} \hat{\mathbf{G}}_m^\dagger + \boldsymbol{\Omega}_m\right) \mathbf{B}_m^{0\dagger}\right] \\
&\quad + \text{tr}\left[\mathbf{A}_m^0 \overline{\boldsymbol{\Gamma}}_{m-1}^0 \mathcal{S}_{m-1} \overline{\boldsymbol{\Gamma}}_{m-1}^{0\dagger} \mathbf{A}_m^{0\dagger}\right] \\
&\quad + \text{tr}\left(\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q} \hat{\mathbf{G}}_m^\dagger \mathbf{B}_m^{0\dagger} + h.c.\right) \\
&\quad - \text{tr}\left(\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q} + \mathbf{B}_m^0 \hat{\mathbf{G}}_m \mathbf{Q} + h.c.\right) \\
&\quad + \text{tr}\left(\mathbf{Q}\right) + \text{tr}\left(\boldsymbol{\Sigma}_m\right) \qquad (50)
\end{aligned}$$

where $h.c.$ signifies the hermitian conjugate of the term in the same parenthesis. In the above equation, the second line derives from (22) and (23), while the final equation results from expanding the expression $\mathbb{E}\{\|\bar{\mathbf{s}}_m - \mathbf{q}\|^2\} = \text{tr}[\mathbb{E}\{(\bar{\mathbf{s}}_m - \mathbf{q})(\bar{\mathbf{s}}_m - \mathbf{q})^\dagger\}]$ and rearranging terms. Furthermore, from (18) and (21) follows that $\hat{\mathbf{s}}_m = \mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \bar{\boldsymbol{\upsilon}}_{m-1} + \mathbf{B}_m^0 \bar{\mathbf{y}}_m + \mathbf{A}_m^0 \overline{\boldsymbol{\Gamma}}_{m-1}^0 \boldsymbol{\omega}_{m-1}$ and thus $\text{tr}(\boldsymbol{\Sigma}_m) = \text{tr}(\mathbb{E}\{\hat{\mathbf{s}}_m \hat{\mathbf{s}}_m^\dagger\})\epsilon$ also depends on $\mathbf{A}_m^0$ and $\mathbf{B}_m^0$. Going through the same algebra as above, we find that $\text{tr}(\mathbb{E}\{\hat{\mathbf{s}}_m \hat{\mathbf{s}}_m^\dagger\})$ is simply the first five lines of the last equality in (50). Hence, the MSE_m is reformatted as

$$\begin{aligned}
\text{MSE}_m &= \text{tr}\left(\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q} \hat{\mathbf{H}}_{m-1}^\dagger \overline{\mathbf{B}}_{m-1}^{0\dagger} \mathbf{A}_m^{0\dagger}\right) \gamma_\epsilon^{-1} \\
&\quad + \text{tr}\left(\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \mathcal{K}_{m-1} \overline{\mathbf{B}}_{m-1}^{0\dagger} \mathbf{A}_m^{0\dagger}\right) \gamma_\epsilon^{-1} \\
&\quad + \text{tr}\left[\mathbf{B}_m^0 \left(\hat{\mathbf{G}}_m \mathbf{Q} \hat{\mathbf{G}}_m^\dagger + \boldsymbol{\Omega}_m\right) \mathbf{B}_m^{0\dagger}\right] \gamma_\epsilon^{-1} \\
&\quad + \text{tr}\left(\mathbf{A}_m^0 \overline{\boldsymbol{\Gamma}}_{m-1}^0 \mathcal{S}_{m-1} \overline{\boldsymbol{\Gamma}}_{m-1}^{0\dagger} \mathbf{A}_m^{0\dagger}\right) \gamma_\epsilon^{-1} \\
&\quad + \text{tr}\left(\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q} \hat{\mathbf{G}}_m^\dagger \mathbf{B}_m^{0\dagger} + h.c.\right) \gamma_\epsilon^{-1} \\
&\quad - \text{tr}\left(\mathbf{A}_m^0 \overline{\mathbf{B}}_{m-1}^0 \hat{\mathbf{H}}_{m-1} \mathbf{Q} + \mathbf{B}_m^0 \hat{\mathbf{G}}_m \mathbf{Q} + h.c.\right) \\
&\quad + \text{tr}\left(\mathbf{Q}\right) \qquad (51)
\end{aligned}$$

where $\gamma_\epsilon^{-1} = 1+\epsilon$. Since (51) contains quadratic expressions of $\mathbf{B}_m^0$ and $\mathbf{A}_m^0$, the MSE_m can readily be minimized by setting the gradient, with respect to the elements of $\mathbf{B}_m^0$ and $\mathbf{A}_m^0$, equal to zero. Defining the matrices $\nabla_{\mathbf{A}}\mathrm{MSE}_m$ and $\nabla_{\mathbf{B}}\mathrm{MSE}_m$ as $[\nabla_{\mathbf{A}}\mathrm{MSE}_m]_{i,j} = \frac{\partial \mathrm{MSE}_m}{\partial [\mathbf{A}_m^0]_{i,j}}$ and $\nabla_{\mathbf{B}}\mathrm{MSE}_m = \frac{\partial \mathrm{MSE}_m}{\partial [\mathbf{B}_m^0]_{i,a}}$, for $i,j \in [K]$ and $a \in [L]$, we have

$$\begin{aligned}\nabla_{\mathbf{A}}\mathrm{MSE}_m = 0 \\ \Rightarrow \gamma_\epsilon^{-1}\mathbf{A}_m^0\overline{\mathbf{B}}_{m-1}^0\left(\hat{\mathbf{H}}_{m-1}\mathbf{Q}\hat{\mathbf{H}}_{m-1}^\dagger + \mathcal{K}_{m-1}\right)\overline{\mathbf{B}}_{m-1}^{0\dagger} \\ + \gamma_\epsilon^{-1}\mathbf{A}_m^0\overline{\mathbf{\Gamma}}_{m-1}^0\mathcal{S}_{m-1}\overline{\mathbf{\Gamma}}_{m-1}^{0\dagger} - \mathbf{Q}\hat{\mathbf{H}}_{m-1}^\dagger\overline{\mathbf{B}}_{m-1}^{0\dagger} \\ + \gamma_\epsilon^{-1}\mathbf{B}_m^0\hat{\mathbf{G}}_m\mathbf{Q}\hat{\mathbf{H}}_{m-1}^\dagger\overline{\mathbf{B}}_{m-1}^{0\dagger} = 0 \end{aligned} \tag{52}$$

and

$$\begin{aligned}\nabla_{\mathbf{B}}\mathrm{MSE}_m = 0 \\ \Rightarrow \gamma_\epsilon^{-1}\mathbf{A}_m^0\overline{\mathbf{B}}_{m-1}^0\hat{\mathbf{H}}_{m-1}\mathbf{Q}\hat{\mathbf{G}}_m^\dagger \\ + \gamma_\epsilon^{-1}\mathbf{B}_m^0\hat{\mathbf{G}}_m\mathbf{Q}\hat{\mathbf{G}}_m^\dagger \\ + \gamma_\epsilon^{-1}\mathbf{B}_m^0\mathbf{\Omega}_m - \mathbf{Q}\hat{\mathbf{G}}_m^\dagger = 0 \end{aligned} \tag{53}$$

Solving the above linear system of equations for the matrices $\mathbf{A}_m^0$ and $\mathbf{B}_m^0$, we conclude to the expressions given in (24) and (25).

Furthermore, in order for each APU m to calculate the $\mathbf{B}_m^0$ and $\mathbf{A}_m^0$, it needs to have knowledge over local information as well as over the matrices $\mathbf{F}_{m-1}^0$ and $\overline{\mathbf{B}}_{m-1}^0\hat{\mathbf{H}}_{m-1}\mathbf{Q}$ that has received from APU $m-1$. Then, according to (22), (23) and (26), it can evaluate these matrices as

$$\begin{aligned}\mathbf{F}_m^0 = \mathbf{A}_m^0\mathbf{F}_{m-1}^0\mathbf{A}_m^{0\dagger} + \mathbf{A}_m^0\overline{\mathbf{B}}_{m-1}^0\hat{\mathbf{H}}_{m-1}\mathbf{Q}\hat{\mathbf{G}}_m^\dagger\mathbf{B}_m^{0\dagger} \\ + \mathbf{B}_m^0\left(\hat{\mathbf{G}}_m\mathbf{Q}\hat{\mathbf{G}}_m^\dagger + \mathbf{\Omega}_m\right)\mathbf{B}_m^{0\dagger} + \mathbf{\Sigma}_m \\ + \mathbf{B}_m^0\hat{\mathbf{G}}_m\mathbf{Q}\hat{\mathbf{H}}_{m-1}^\dagger\overline{\mathbf{B}}_{m-1}^{0\dagger}\mathbf{A}_m^{0\dagger} \end{aligned} \tag{54}$$

and

$$\overline{\mathbf{B}}_m^0\hat{\mathbf{H}}_m\mathbf{Q} = \mathbf{A}_m^0\overline{\mathbf{B}}_{m-1}^0\hat{\mathbf{H}}_{m-1}\mathbf{Q} + \mathbf{B}_m^0\hat{\mathbf{G}}_m\mathbf{Q} \tag{55}$$

Appendix B
Proof Of Hermitian Matrices

Here we show that the matrices $\mathbf{F}_m^0$ and $\overline{\mathbf{B}}_m^0\hat{\mathbf{H}}_m\mathbf{Q}$, which are exchanged between the APUs as side information, are Hermitian and thus they can be represented by $2K^2$ real-valued symbols. From (26), we observe that $\mathbf{F}_m^0 = \mathbf{F}_m^{0\dagger}$, thus making the $\mathbf{J}_m$ and $\mathbf{\Lambda}_m$ Hermitian according to (27) and (28), respectively. By analyzing (55), $\overline{\mathbf{B}}_m^0\hat{\mathbf{H}}_m\mathbf{Q}$ becomes

$$\begin{aligned}\overline{\mathbf{B}}_m^0\hat{\mathbf{H}}_m\mathbf{Q} &= \mathbf{A}_m^0\overline{\mathbf{B}}_{m-1}^0\hat{\mathbf{H}}_{m-1}\mathbf{Q} + \mathbf{B}_m^0\hat{\mathbf{G}}_m\mathbf{Q} \\ &\overset{(25)}{=} \left(\gamma_\epsilon\mathbf{I}_K - \mathbf{B}_m^0\hat{\mathbf{G}}_m\right)\mathbf{\Lambda}_{m-1} + \mathbf{B}_m^0\hat{\mathbf{G}}_m\mathbf{Q} \\ &= \gamma_\epsilon\mathbf{\Lambda}_{m-1} + \mathbf{B}_m^0\hat{\mathbf{G}}_m\left(\mathbf{Q} - \mathbf{\Lambda}_{m-1}\right) \\ &\overset{(24)}{=} \gamma_\epsilon(\mathbf{Q} - \mathbf{\Lambda}_{m-1})\,\hat{\mathbf{G}}_m^\dagger\mathbf{J}_m^{-1}\hat{\mathbf{G}}_m\left(\mathbf{Q} - \mathbf{\Lambda}_{m-1}\right) \\ &\quad + \gamma_\epsilon\mathbf{\Lambda}_{m-1} \end{aligned} \tag{56}$$

Considering the above equation and due to the fact that $\mathbf{\Lambda}_{m-1}$ and $\mathbf{J}_m^{-1}$ are Hermitian matrices, it derives that $\overline{\mathbf{B}}_m^0\hat{\mathbf{H}}_m\mathbf{Q}$ is also Hermitian.

References

[1] I. Chiotis and A. L. Moustakas, "On the uplink performance of finite-capacity radio stripes," in *Proc. IEEE Int. Medit. Conf. Commun. Netw. (MeditCom)*, Sep. 2022, pp. 118–123.

[2] I. Chiotis and A. L. Moustakas, "Optimal MMSE processing for limited-capacity radio stripes," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2023, pp. 1–6.

[3] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[4] 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.

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

[6] 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.

[7] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, Aug. 2019, Art. no. 197.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] J. Kassam, D. Castanheira, A. Silva, R. Dinis, and A. Gameiro, "Joint decoding and UE-APs association for scalable cell-free systems," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7303–7315, Dec. 2023.

[14] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[15] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "User-centric cell-free massive MIMO networks: A survey of opportunities, challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 611–652, 1st Quart., 2022.

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

[17] 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.

[18] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[19] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, Feb. 2020.

[20] I. Atzeni, B. Gouda, and A. Tölli, "Distributed precoding design via over-the-air signaling for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1201–1216, Feb. 2021.

[21] Z. H. Shaik, E. Björnson, and E. G. Larsson, "MMSE-optimal sequential processing for cell-free massive MIMO with radio stripes," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7775–7789, Nov. 2021.

[22] Z. H. Shaik, E. Björnson, and E. G. Larsson, "Cell-free massive MIMO with radio stripes and sequential uplink processing," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.

[23] P. Frenger, J. Hederen, M. Hessler, and G. Interdonato, "Improved antenna arrangement for distributed massive MIMO," WO Patent 2 018 103 897 A1, Jun. 14, 2018.

[24] P. Zhang and F. M. J. Willems, "On the downlink capacity of cell-free massive MIMO with constrained fronthaul capacity," *Entropy*, vol. 22, no. 4, p. 418, Apr. 2020. [Online]. Available: https://www.mdpi.com/1099-4300/22/4/418

[25] O. L. A. López et al., "Energy-sustainable IoT connectivity: Vision, technological enablers, challenges, and future directions," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 2609–2666, 2023.

[26] A. Fascista et al., "Uplink joint positioning and synchronization in cell-free deployments with radio stripes," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2023, pp. 1330–1336.

[27] O. L. A. López, D. Kumar, R. D. Souza, P. Popovski, A. Tölli, and M. Latva-Aho, "Massive MIMO with radio stripes for indoor wireless energy transfer," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7088–7104, Sep. 2022.

[28] F. Conceição, L. Martins, M. Gomes, V. Silva, and R. Dinis, "Access point selection for spectral efficiency and load balancing optimization in radio stripes," *IEEE Commun. Lett.*, vol. 27, no. 9, pp. 2383–2387, Sep. 2023.

[29] H. Li, Y. Dong, C. Gong, X. Wang, and X. Dai, "Gaussian message passing detection with constant front-haul signaling for cell-free massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5395–5400, Apr. 2023.

[30] Ó. Martins, F. Conceição, M. Gomes, V. Silva, and R. Dinis, "Achievable capacity for continuous radio stripe LOS communications," *IEEE Commun. Lett.*, vol. 27, no. 10, pp. 2792–2796, Oct. 2023.

[31] A. M. Sonagara, M. Mishra, R. S. Kshetrimayum, E. Bjornson, and Z. N. Chen, "Ultra-thin flexible uniplanar antenna based on SSPP for B5G radio stripe network," *IEEE Antennas Wireless Propag. Lett.*, vol. 22, no. 8, pp. 1947–1951, Aug. 2023.

[32] J. Kaleva, A. Tölli, M. Juntti, R. A. Berry, and M. L. Honig, "Decentralized joint precoding with pilot-aided beamformer estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2330–2341, May 2018.

[33] E. Shi, J. Zhang, J. Zhang, D. W. K. Ng, and B. Ai, "Decentralized coordinated precoding design in cell-free massive MIMO systems for URLLC," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2638–2642, Feb. 2023.

[34] E. Björnson and L. Sanguinetti, "A new look at cell-free massive MIMO: Making it practical with dynamic cooperation," in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 1–6.

[35] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Cell-free massive MIMO with large-scale fading decoding and dynamic cooperation clustering," in *Proc. 24th Int. ITG Workshop Smart Antennas (WSA)*, Nov. 2021, pp. 1–6.

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

[37] C. Hao, T. T. Vu, H. Q. Ngo, M. N. Dao, X. Dang, and M. Matthaiou, "User association and power control in cell-free massive MIMO with the APG method," in *Proc. 31st Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2023, pp. 1469–1473.

[38] G. Femenias, F. Riera-Palou, and E. Björnson, "Another twist to the scalability in cell-free massive MIMO networks," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6793–6804, Nov. 2023.

[39] Y. Ma, Z. Yuan, G. Yu, and Y. Chen, "Cooperative scheme for cell-free massive MIMO with radio stripes," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2021, pp. 1–6.

[40] Z. Zhang, Y. Dong, K. Long, X. Wang, and X. Dai, "Decentralized baseband processing with Gaussian message passing detection for uplink massive MU-MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 2152–2157, Feb. 2022.

[41] S. Park, O. Simeone, O. Sahin, and S. S. Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[42] 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.

[43] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, and M. Debbah, "Cell-free massive MIMO with limited backhaul," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[44] 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.

[45] M. Bashar, H. Q. Ngo, A. G. Burr, D. Maryopi, K. Cumanan, and E. G. Larsson, "On the performance of backhaul constrained cell-free massive MIMO with linear receivers," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 624–628.

[46] 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.

[47] M. Bashar et al., "Uplink spectral and energy efficiency of cell-free massive MIMO with optimal uniform quantization," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 223–245, Jan. 2021.

[48] B. Sklar and P. K. Ray, *Digital Communications: Fundamentals and Applications*, 2nd ed. London, U.K.: Pearson Education, 2008.

[49] S. Jo, H. Lee, and S.-H. Park, "Joint precoding and fronthaul compression for cell-free MIMO downlink with radio stripes," 2023, *arXiv:2308.03251*.

[50] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing), 2nd ed. Hoboken, NJ, USA: Wiley, Jul. 2006.

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

[52] M. Fiedler, "Bounds for the determinant of the sum of Hermitian matrices," *Proc. Amer. Math. Soc.*, vol. 30, no. 1, pp. 27–31, 1971.

[53] *Further Advancements for E-ULTRA Physical Layer Ascpects (Release 9)*, document TS 36.814, 3GPP, Mar. 2017.

Ioannis Chiotis (Graduate Student Member, IEEE) received the Electronics and Telecommunications degree from the Hellenic Air Force Academy, Greece, in 2015, and the master's degree in radio-electrology and electronics from the Department of Physics, National and Kapodistrian University of Athens (NKUA), Greece, in 2021, where he is currently pursuing the Ph.D. degree. His research interests include cell-free massive MIMO systems, signal processing for communications, and convex optimization.

Aris L. Moustakas (Senior Member, IEEE) received the B.S. degree in physics from Caltech and the M.S. degree in physics and the Ph.D. degree in theoretical condensed matter physics from Harvard University. He worked with Bell Labs, Lucent Technologies, USA, from 1998 to 2005, first as a Post-Doctoral Researcher and then as a Member of Technical Staff. He then joined the Faculty of the Physics Department, National and Kapodistrian University of Athens. He has held the Senior DIGITEO Chair in Paris, France, from 2013 to 2014. He is currently the Lead Researcher with the Archimedes/Athena Research Center, Athens. His main research interests include multiple antenna systems and the applications of game theory and statistical physics to communications and networks and more recently machine learning. He served as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION THEORY from 2009 to 2012.