Sequential Processing in Cell-free Massive MIMO Uplink with Limited Memory Access Points

Vida Ranjbar*, Robbert Beerten*, Marc Moonen*, Sofie Pollin[†]

*Department of Electrical Engineering, KU Leuven, Belgium
[†]IMEC, Kapeldreef 75, 3001 Leuven, Belgium
Corresponding Author: {vida.ranjbar}@kuleuven.be

Abstract—Cell-free massive multiple-input multiple-output (MIMO) is an emerging technology that will reshape the architecture of next-generation networks. This paper considers the sequential fronthaul, whereby the access points (APs) are connected in a daisy chain topology with multiple sequential processing stages. With this sequential processing in the uplink, each AP refines users' signal estimates received from the previous AP based on its own local received signal vector. While this processing architecture has been shown to achieve the same performance as centralized processing, the impact of the limited memory capacity at the APs on the store and forward processing architecture is yet to be analyzed. Thus, we model the received signal vector compression using rate-distortion theory to demonstrate the effect of limited memory capacity on the optimal number of APs in the daisy chain fronthaul. Without this memory constraint, more geographically distributed antennas alleviate the adverse effect of large-scale fading on the signal-to-interference-plus-noise-ratio (SINR). However, we show that in case of limited memory capacity at each AP, the memory capacity to store the received signal vectors at the final AP of this fronthaul becomes a limiting factor. In other words, we show that when deciding on the number of APs to distribute the antennas, there is an inherent trade-off between more macro-diversity and compression noise power on the stored signal vectors at the APs. Hence, the available memory capacity at the APs significantly influences the optimal number of APs in the fronthaul.

Index Terms—Uplink cell-free massive MIMO network, daisy chain fronthaul topology, sequential processing, limited memory capacity constraint, macro diversity.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is one of the critical enablers for next-generation mobile networks. It promises spectral efficiency (SE), energy efficiency (EE), and reliability and allows for low-cost hardware at both receiver and transmitter [1]. Massive MIMO provides additional degrees of freedom in the spatial domain, allowing it to separate users spatially rather than via time and frequency scheduling. This reuse of time and frequency resources dramatically increases the average throughput. Cell-free massive MIMO has attracted a lot of academic attention recently as a promising massive MIMO technology for the future generation of wireless networks due to its ability to mitigate the adverse effect of large-scale fading on the users' signal-to-interference-plus-noise-ratio (SINR) and to provide uniform service to all users [2]. In such a network, the antennas are distributed among the access points (APs), and all or a few nearby APs will serve each user. As the serving APs are selected based on

their vicinity to the user, the cell-edge phenomenon and poor coverage of traditional cellular networks disappear.

In cell-free massive MIMO networks, the APs cooperate to serve the users with direct information exchange, e.g., in a sequential fronthaul topology [3]–[6] or indirectly through a central processing unit (CPU) in a star fronthaul topology [2], [7], [8]. In the uplink of a cell-free massive MIMO network with a daisy chain fronthaul topology, each AP estimates the users' signal and sends the local estimates to the next AP in the sequence. In this way, the users' signal estimates are refined through the daisy chain fronthaul. Hence, sequential processing in a daisy chain fronthaul topology requires the AP to store their received signal vector in the memory until they receive the corresponding information from the previous AP in the sequence. In [4], it is proven that with sequential processing at each AP connected in a daisy chain fronthaul topology, and for the same number of exchanged scalars on each chunk of fronthaul, the minimum mean square error (MMSE) optimal solution can be achieved in the last AP. However, the memory constraint at each AP in the sequential fronthaul is neglected.

Non-idealities such as limited bandwidth fronthaul links, hardware impairment, and low-resolution analog-to-digital converters (ADCs) in both cellular and cell-free massive MIMO networks are discussed in [9]–[16], among others. In [9], it is shown under which scenarios the correlation between the distortion vector elements in a massive MIMO network with hardware impairment has a negligible impact on the users' SE. The ADC bit allocation among antennas in a cell-free massive MIMO network is discussed in [10]–[13]. In [11], the SE and EE maximization problems are formulated as a function of the number of ADC bits used to represent the antenna signals, subject to a constraint on the total number of bits or power consumption. In [13], adaptive intra-AP and inter-AP bit allocation for ADCs is considered. In [14], [15], the impact of limited capacity fronthaul links on the users' SE and EE in cell-free massive MIMO uplink is investigated.

The number of bits to quantize the received signal vector is a vital cost metric for the ADC, in the case of limited fronthaul capacity and when the vector needs to be stored in memory in a network with sequential fronthaul topology, such as in this paper. The sequential fronthaul topology is studied in, e.g. [3]–[6]. However, these works do not consider the limited capacity of the memory at the APs. This paper studies

sequential processing for uplink users' signal estimation in a cell-free massive MIMO network with a daisy chain fronthaul topology under the realistic assumption of a limited memory capacity constraint at each AP.

Contribution: To the best of the authors' knowledge, this paper is the first to address the problem of limited memory capacity availability at the APs due to the sequential processing in the cell-free massive MIMO with daisy chain fronthaul. First, we use a tractable model based on rate-distortion theory to model the compression of the received signal vectors in the memory of the APs. Second, using two limited memory capacity models, the effect of limited memory capacity on the optimal number of APs in the daisy chain fronthaul topology is quantified in the simulation section.

A. Notation

We denote vectors and matrices with boldface lower-case and upper-case letters, respectively. Transpose and conjugate transpose operations are denoted by superscripts $^{\mathrm{T}}$ and $^{\mathrm{H}}$, respectively. A circularly symmetric complex Gaussian distribution with covariance matrix $\mathbf{X}$ is represented as $\mathcal{CN}(0, \mathbf{X})$. Symbol $\mathbb{E}\{\mathbf{x}\}$ denotes the mean of $\mathbf{x}$. $\mathcal{H}(\mathbf{x})$ is the differential entropy of $\mathbf{x}$, and $I(\mathbf{x}; \hat{\mathbf{x}})$ is the mutual information between $\mathbf{x}$ and $\hat{\mathbf{x}}$. The Euclidean norm of $\mathbf{x}$ is shown as $\|\mathbf{x}\|$. We use $\mathrm{diag}(\mathbf{X})$ to signify the elements on the main diagonal of $\mathbf{X}$ and $\mathrm{diag}(\mathbf{x})$ for a diagonal matrix with $\mathbf{x}$ as its main diagonal. Furthermore, $\mathbf{X} = \mathrm{blkdiag}(\mathbf{X}_1, \ldots, \mathbf{X}_L)$ is a block-diagonal matrix with matrices $\mathbf{X}_i \; i = \{1, \ldots, L\}$ as diagonal blocks. $\mathbf{X}^{1/2}$ is the square-root of $\mathbf{X}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semi-definite. Finally, $\mathrm{tr}(\mathbf{X})$ denotes the trace of matrix $\mathbf{X}$.

II. SYSTEM MODEL AND PROBLEM STATEMENT

Distributed processing in cell-free massive MIMO networks is a necessity as it avoids overloading a single AP with massive computations, and it enables truly scalable implementations and, hence, large-scale deployments. Distributed processing has become even more attractive because of the growing interest in the sequential fronthaul topology [3]–[6]. In this paper, we consider distributed uplink signal estimation using the least-squares (LS) method in a cell-free massive MIMO network with limited memory APs. There are L APs, each having N antennas, connected in a daisy chain fronthaul topology, serving K single antenna users.

A. Recursive least-squares (RLS) for uplink signal estimation

The received signal vector at AP l in the uplink is given as follows:

$$\mathbf{y}_l = \mathbf{H}_l \mathbf{s} + \mathbf{n}_l, \tag{1}$$

where $\mathbf{s} \sim \mathcal{CN}(0, p\mathbf{I}_K)$ is the users' signal, $\mathbf{H}_l \in \mathbb{C}^{N \times K}$ and $\mathbf{n}_l \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ are the local channel matrix and noise vector at AP l, respectively. The channel vector between user k and AP l is drawn from a correlated Rayleigh distribution, i.e. $\mathbf{H}_{l[:,k]} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{kl})$, where subscript $[:, k]$ denotes the kth column of $\mathbf{H}_l$ and $\beta_{kl} = \mathrm{tr}(\mathbf{R}_{kl})/N$. We assume a block fading model in which the channel matrix $\mathbf{H}_l$ remains constant

in a coherence interval of $\tau_c = B_c T_c$ samples, with T_c and B_c the coherence time and coherence bandwidth of the channel, respectively [17]. Out of τ_c samples, τ_u samples are used for the uplink. We assume perfect channel state information (CSI) at the APs, which is possible with a unique pilot per user and high enough transmission power during pilot transmission. A compressed version of the received vector $\hat{\mathbf{y}}_l$ is stored in the local memory of AP l and is defined as follows:

$$\hat{\mathbf{y}}_l = \mathbf{y}_l + \mathbf{q}_l = \mathbf{H}_l \mathbf{s} + \mathbf{z}_l, \tag{2}$$

where $\mathbf{z}_l = \mathbf{n}_l + \mathbf{q}_l$ is a spatially correlated noise vector with zero mean and covariance matrix $\mathbf{Z}_l$. The network-wide compressed received signal and noise vector can be expressed as $\hat{\mathbf{y}} = \begin{bmatrix} \hat{\mathbf{y}}_1^{\mathrm{T}} & \ldots & \hat{\mathbf{y}}_L^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ and $\mathbf{z} = \begin{bmatrix} \mathbf{z}_1^{\mathrm{T}} & \ldots & \mathbf{z}_L^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$, respectively. The noise vectors in different APs are assumed to be independent, i.e., $\mathbf{Z} = \mathrm{blkdiag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_L)$.

Algorithm 1 summarizes the RLS steps for sequential uplink signal estimation among APs [6]. Note that the superscript n in algorithm 1 differentiates the uplink samples in one coherence block. However, this superscript is not used anywhere else in the paper for notational simplicity. By updating the estimates

---

Algorithm 1 RLS algorithm for users' signal estimation

1: **Initialize:**
2:     $\mathbf{\Gamma}_0 = p\mathbf{I}_K$
3:     $\hat{\mathbf{s}}_0^n = \mathbf{0}_{K \times 1}, \forall n \in [1 : \tau_u]$
4: **for** $l = 1 \ldots L$ **do**
5:     $\mathbf{\Gamma}_l = \mathbf{\Gamma}_{l-1} - \mathbf{\Gamma}_{l-1}\mathbf{H}_l^{\mathrm{H}}\mathbf{Z}_l^{-\mathrm{H}/2}(\mathbf{I}_N + \mathbf{Z}_l^{-1/2}\mathbf{H}_l\mathbf{\Gamma}_{l-1}\mathbf{H}_l^{\mathrm{H}}\mathbf{Z}_l^{-\mathrm{H}/2})^{-1}\mathbf{Z}_l^{-1/2}\mathbf{H}_l\mathbf{\Gamma}_{l-1}^{\mathrm{H}}$
6:     **for** $n = 1 \ldots \tau_u$ **do**
7:         $\hat{\mathbf{s}}_l^n = \hat{\mathbf{s}}_{l-1}^n + \mathbf{\Gamma}_l\mathbf{H}_l^{\mathrm{H}}\mathbf{Z}_l^{-\mathrm{H}/2}(\hat{\mathbf{y}}_l^n - \mathbf{Z}_l^{-1/2}\mathbf{H}_l\hat{\mathbf{s}}_{l-1}^n)$.
8:     **end for**
9: **end for**

---

of the users' signal as in algorithm 1, the users' signal estimates at the final AP will be:

$$\hat{\mathbf{s}} = (\mathbf{H}^{\mathrm{H}}\mathbf{Z}^{-1}\mathbf{H} + \frac{1}{p}\mathbf{I}_k)^{-1}\mathbf{H}^{\mathrm{H}}\mathbf{Z}^{-1}\hat{\mathbf{y}}. \tag{3}$$

In a sequential fronthaul, the number of received signal vectors to be stored grows linearly with the number of APs in the network and should be processed at every symbol time. Thus, this memory should be very fast so that every sample can be processed promptly. On the contrary, the local CSI should be stored only once per coherence block in each AP. Section II-B elaborates on storing the local received signal vectors in the limited memory. In Section III, the compression of the local received signal vector at each AP is defined from a sum-SE optimization problem constrained by the limited memory capacity at the AP.

B. Storage of received signal vectors in the limited memory APs

In a daisy chain fronthaul, the APs estimate each user's signal based on their local received signal vector and the signal estimates they receive from the previous AP, as shown in line

Fig. 1. Sequential processing and storage in a daisy chain fronthaul topology

(7) of algorithm 1. Therefore, they need to store their local received signal vector until the previous AP has finished processing its corresponding local received signal vector. To make the problem more tangible, consider an orthogonal frequency-division multiplexing (OFDM) system where the bandwidth is divided between F subcarriers. Each AP receives N OFDM symbols, one for each antenna. Suppose that $\mathbf{Y}_l^{t_0} \in \mathbb{C}^{N \times F}$ is a symbol matrix with each column corresponding to the received signal vector at AP l for a particular subcarrier of symbol t_0, as shown in Fig 1. When AP 1 is processing $\mathbf{Y}_1^{t_0}$, the rest of the APs have their corresponding matrices, i.e., $\mathbf{Y}_l^{t_0}, \forall l \in \{2, \dots, L\}$ in their memory. Similarly, when AP 2 is processing $\mathbf{Y}_2^{t_0-1}$, the corresponding matrix at AP l, i.e., $\mathbf{Y}_l^{t_0-1}, \forall l \in \{3, \dots, L\}$, is stored in the memory. Accordingly, the number of the symbol matrix stored at AP l is $l-1$ meaning that the number of received signal vectors stored at AP l is $(l-1)F$, which increases linearly from one AP to the next by F. It is worth mentioning that processors are designed to process at least one symbol during a symbol duration to have a stable system. In other words, the rate of the locally received signal vectors entering the memory should be lower or, in the worst case, the same as the rate at which the processor is processing them. Therefore, the number of vectors stored in the AP's memory increases by F from one AP to the next.

The memory to store the vectors is usually a fast on-chip cache memory [18], close to the processing unit [19].

III. RECEIVED SIGNAL VECTOR COMPRESSION

In this section, for storing the received signal vectors, we consider 1) Joint compression of the received signal vector elements, also called vector-wise compression (VC) of the received signal vector, and 2) Element-wise compression (EC) of the received signal vector.

A. Option 1: Vector-wise compression (VC) of the received signal vector

The compressed vector at AP l can be represented as $\hat{\mathbf{y}}_{vl}$ and the relation between $\mathbf{y}_l$ and $\hat{\mathbf{y}}_{vl}$ follows as [20], [21]:

$$\hat{\mathbf{y}}_{vl} = \mathbf{y}_l + \mathbf{q}_{vl} = \mathbf{H}_l \mathbf{s} + \mathbf{n}_l + \mathbf{q}_{vl}, \qquad (4)$$

where $\mathbf{q}_{vl} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_{vl})$ represents the compression noise which is independent of $\mathbf{y}_l$. We also define $\mathbf{z}_{vl} = \mathbf{n}_l + \mathbf{q}_{vl}$, with covariance matrix $\mathbf{Z}_{vl} = \mathbb{E}\{\mathbf{z}_{vl}\mathbf{z}_{vl}^H\} = \mathbf{Q}_{vl} + \sigma^2 \mathbf{I}_N$.

The relation between the number of bits to compress the received signal vector, C_s, and the compression noise covari-

ance matrix $\mathbf{Q}_{vl}$ at AP l, conditioned on the local CSI is as follows:

$$
\begin{aligned}
C_s &= I(\mathbf{y}_l; \hat{\mathbf{y}}_{vl} | \mathbf{H}_l) \\
&= \mathcal{H}(\hat{\mathbf{y}}_{vl} | \mathbf{H}_l) - \mathcal{H}(\hat{\mathbf{y}}_{vl} | \mathbf{y}_l, \mathbf{H}_l) \qquad (5) \\
&= \log_2 \det(\mathbf{Q}_{vl}^{-1}(p\mathbf{H}_l\mathbf{H}_l^H + \sigma^2 \mathbf{I}_N) + \mathbf{I}_N).
\end{aligned}
$$

The network-wide compressed received vector is given as:

$$\hat{\mathbf{y}}_v = \mathbf{y} + \mathbf{q}_v = \mathbf{H}\mathbf{s} + \underbrace{\mathbf{n} + \mathbf{q}_v}_{\mathbf{z}_v}, \qquad (6)$$

where $\hat{\mathbf{y}}_v = \begin{bmatrix} \hat{\mathbf{y}}_{v1}^T & \cdots & \hat{\mathbf{y}}_{vL}^T \end{bmatrix}^T$, $\mathbf{q}_v = \begin{bmatrix} \mathbf{q}_{v1}^T & \cdots & \mathbf{q}_{vL}^T \end{bmatrix}^T$, and $\mathbf{z}_v = \begin{bmatrix} \mathbf{z}_{v1}^T & \cdots & \mathbf{z}_{vL}^T \end{bmatrix}^T$ is the receiver plus compression noise vector with covariance matrix $\mathbf{Z}_v = \mathbb{E}\{\mathbf{z}_v\mathbf{z}_v^H\} = \text{blkdiag}(\mathbf{Z}_{v1}, \dots, \mathbf{Z}_{vL}) = \text{blkdiag}(\mathbf{Q}_{v1} + \sigma^2 \mathbf{I}_N, \dots, \mathbf{Q}_{vL} + \sigma^2 \mathbf{I}_N)$. Following the discussion in Section II-A, the LS estimates of the users' signal can be formulated as below:

$$\hat{\mathbf{s}}_v = \mathbf{V}_v \hat{\mathbf{y}}_v, \qquad (7)$$

where combiner matrix $\mathbf{V}_v$ is given as follows:

$$\mathbf{V}_v = (\mathbf{H}^H \mathbf{Z}_v^{-1} \mathbf{H} + \frac{1}{p}\mathbf{I}_k)^{-1} \mathbf{H}^H \mathbf{Z}_v^{-1}. \qquad (8)$$

Having the estimates of the users' signal as in (7), the sum-SE of users is formulated as follows:

$$
\begin{aligned}
R_{VC} &= \frac{\tau_u}{\tau_c} \mathcal{I}(\mathbf{V}_v \hat{\mathbf{y}}; \mathbf{s} | \mathbf{V}_v, \mathbf{H}) \\
&= \frac{\tau_u}{\tau_c} \Big(\mathcal{H}(\mathbf{V}_v \hat{\mathbf{y}}_v | \mathbf{V}_v, \mathbf{H}) - \mathcal{H}(\mathbf{V}_v \hat{\mathbf{y}}_v | \mathbf{s}, \mathbf{V}_v, \mathbf{H}) \Big) \\
&= \frac{\tau_u}{\tau_c} \log_2 \det(p\mathbf{H}\mathbf{H}^H \mathbf{Z}_v^{-1} + \mathbf{I}_{NL}) \\
&\overset{(a)}{\leq} \frac{\tau_u}{\tau_c} \log_2 \prod_{l=1}^{L} \det(p\mathbf{H}_l\mathbf{H}_l^H \mathbf{Z}_{vl}^{-1} + \mathbf{I}_N) \qquad (9) \\
&= \frac{\tau_u}{\tau_c} \sum_{l=1}^{L} \log_2 \det(p\mathbf{H}_l\mathbf{H}_l^H \mathbf{Z}_{vl}^{-1} + \mathbf{I}_N) \\
&= \frac{\tau_u}{\tau_c} \sum_{l=1}^{L} \log_2 \det(p\mathbf{H}_l\mathbf{H}_l^H (\mathbf{Q}_{vl} + \sigma^2 \mathbf{I}_N)^{-1} + \mathbf{I}_N),
\end{aligned}
$$

(a)

where the inequality $\leq$ holds due to the fact that matrix $\mathbf{Z}_v$ is a block-diagonal matrix, and the fact that the determinant of the positive (semi-) definite matrix is always smaller than the determinant of a diagonal matrix with the same diagonal elements [22]. The detailed proof is omitted due to space limitations. The upper bound in equation (9) is for the instantaneous sum-SE in a particular coherence block.

A maximization problem can be formulated to find the optimal $\mathbf{Q}_{vl}, \forall l$ that maximizes the upper bound defined in (9).

The upper bound on the sum-SE in (9) is a summation of L functions each of which depends only on the compression plus receiver noise covariance matrix of a single AP. Additionally, each AP compresses its received signal vector in isolation from other APs. Therefore, maximization of the upper bound can be decomposed into L smaller optimization problems to be solved

at L APs. Accordingly, the sum-SE maximization problem at AP l is defined as follows:

$$\arg\max_{\mathbf{Q}_{vl}^{-1} \succeq 0} \quad \log_2 \det(p\mathbf{H}_l\mathbf{H}_l^{\mathsf{H}}(\mathbf{Q}_{vl} + \sigma^2\mathbf{I}_N)^{-1} + \mathbf{I}_N)$$
$$\text{s.t.} \quad C_s = \log_2 \det(\mathbf{Q}_{vl}^{-1}(p\mathbf{H}_l\mathbf{H}_l^{\mathsf{H}} + \sigma^2\mathbf{I}_N) + \mathbf{I}_N). \tag{10}$$

Similar to [21] and [23, app. B], the problem in (10) can be converted to an equivalent optimization problem that maximizes the objective function with respect to the eigenvalues of $\mathbf{Q}_{vl}^{-1}$. Due to space limitations, the proof is omitted.

The optimal matrix $\mathbf{Q}_{vl}^{-1}$ is found to be as follows:

$$\mathbf{Q}_{vl}^{-1*} = \mathbf{U}_l \mathbf{\Sigma}_{vlq}^{-1*} \mathbf{U}_l^{\mathsf{H}}, \tag{11}$$

where the columns of $\mathbf{U}_l \in \mathbb{C}^{N \times N}$ are the eigen vectors of $\mathbf{H}_l\mathbf{H}_l^{\mathsf{H}}$ and the ith diagonal element of $\mathbf{\Sigma}_{vlq}^{-1*}$ is found as follows:

$$\lambda_{vlqi}^* = \left[\frac{1}{\mu^*}\left(\frac{1}{\sigma^2} - \frac{1}{p\lambda_{li}^2 + \sigma^2}\right) - \frac{1}{\sigma^2}\right]^+, \forall i, \tag{12}$$

where $\lambda_{li}^2, \forall i \in \{1, \dots, N\}$ are the eigenvalues of $\mathbf{H}_l\mathbf{H}_l^{H}$. Having $\mathbf{Q}_{vl}^{-1*}$, $\mathbf{Q}_{vl}^*$ can be determined accordingly. μ^* is the Lagrange multiplier and is found to satisfy the equality constraint in (10).

B. Option 2: Element-wise compression (EC) of the received signal vector

In the EC of the received vector, each element is compressed individually. The bits allocated to the compression of the ith element of the local received vector $\mathbf{y}_l$ at AP l can be denoted by b_{li} and $C_s = \sum_{i=1}^{N} b_{li}$. The compressed i^{th} element is given as follows:

$$\hat{y}_{eli} = y_{li} + q_{eli} = \mathbf{H}_{l[i,:]}\mathbf{s} + n_{li} + q_{eli}, \tag{13}$$

where the subscript $[i,:]$ specifies the ith row of matrix $\mathbf{H}_l$, $q_{eli} \sim \mathcal{CN}(0, \sigma_{eli}^2)$, $\mathbf{q}_{el} = \begin{bmatrix} q_{el1} & \cdots & q_{elN} \end{bmatrix}^{\mathsf{T}}$ is the compression noise vector with covariance matrix $\mathbf{Q}_{el} = \mathbb{E}\{\mathbf{q}_{el}\mathbf{q}_{el}^{H}\}$. We define $z_{eli} = n_{li} + q_{eli}$, $\mathbf{z}_{el} = \begin{bmatrix} z_{el1} & \cdots & z_{elN} \end{bmatrix}^{\mathsf{T}}$ with covariance matrix $\mathbf{Z}_{el} = \mathbb{E}\{\mathbf{z}_{el}\mathbf{z}_{el}^{H}\} = \mathbf{Q}_{el} + \sigma^2\mathbf{I}_N$. Furthermore, n_{li} is the ith element of the noise vector $\mathbf{n}_l$ and $\hat{\mathbf{y}}_{el} = \begin{bmatrix} \hat{y}_{el1} & \cdots & \hat{y}_{elN} \end{bmatrix}^{\mathsf{T}}$ is the compressed received signal vector. The relation between b_{li} and compression noise of the ith element is:

$$b_{li} = \mathcal{I}(y_{eli}; \hat{y}_{eli}|\mathbf{H}_{l[i,:]})$$
$$= \mathcal{H}(\hat{y}_{eli}|\mathbf{H}_{l[i,:]}) - \mathcal{H}(\hat{y}_{eli}|y_{eli}, \mathbf{H}_{l[i,:]})$$
$$= \log_2\left(\frac{p\|\mathbf{H}_{l[i,:]}\|^2 + \sigma^2}{\sigma_{eli}^2} + 1\right), \tag{14}$$

The diagonal elements of the covariance matrix $\mathbf{Q}_{el}$ can be calculated based on value of $b_{li}, \forall i \in \{1, \dots, N\}$ and represented as $\sigma_{eli}^2, \forall i \in \{1, \dots, N\}$. The off-diagonal elements of matrix $\mathbf{Q}_{el}$ are unknown.

We define diagonal matrix $\mathbf{P}_l$ with the variance of the elements of the local received signal vector $\{\mathbf{y}_l|\mathbf{H}_l\}$ as its diagonal elements,

$$\mathbf{P}_l = \text{diag}(\mathbb{E}\{\mathbf{y}_l\mathbf{y}_l^H|\mathbf{H}_l\})$$
$$= \text{diag}\left(p\|\mathbf{H}_{l[1,:]}\|^2 + \sigma^2, \cdots, p\|\mathbf{H}_{l[N,:]}\|^2 + \sigma^2\right), \tag{15}$$

and the diagonal matrix $\mathbf{Q}_{el}^d$ with variance of the elements of the compression noise vector $\mathbf{q}_{el}$ as its diagonal elements,

$$\mathbf{Q}_{el}^d = \text{diag}\left(\sigma_{el1}^2, \dots, \sigma_{elN}^2\right). \tag{16}$$

We can relate C_s to the variance of the compression noise vector elements as below:

$$C_s = \sum_{i=1}^{N} b_{li} = \log_2 \prod_{i=1}^{N}\left(\frac{p\|\mathbf{H}_{l[i,:]}\|^2 + \sigma^2}{\sigma_{eli}^2} + 1\right)$$
$$= \log_2 \det((\mathbf{Q}_{el}^d)^{-1}\mathbf{P}_l + \mathbf{I}_N). \tag{17}$$

The network-wide compressed received vector is as follows:

$$\hat{\mathbf{y}}_e = \mathbf{y} + \mathbf{q}_e = \mathbf{H}\mathbf{s} + \underbrace{\mathbf{n} + \mathbf{q}_e}_{\mathbf{z}_e}, \tag{18}$$

where $\hat{\mathbf{y}}_e = \begin{bmatrix} \hat{\mathbf{y}}_{e1}^{\mathsf{T}} & \cdots & \hat{\mathbf{y}}_{eL}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$, $\mathbf{q}_e = \begin{bmatrix} \mathbf{q}_{e1}^{\mathsf{T}} & \cdots & \mathbf{q}_{eL}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$ and $\mathbf{z}_e = \begin{bmatrix} \mathbf{z}_{e1}^{\mathsf{T}} & \cdots & \mathbf{z}_{eL}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$ with covariance matrix $\mathbf{Z}_e = \text{blkdiag}(\mathbf{Z}_{e1}, \dots, \mathbf{Z}_{eL}) = \text{blkdiag}(\mathbf{Q}_{e1} + \sigma^2\mathbf{I}_N, \dots, \mathbf{Q}_{eL} + \sigma^2\mathbf{I}_N)$. In EC, the correlation between compression noise elements at one AP, i.e., the off-diagonal elements of $\mathbf{Q}_{el}, \forall l$, are unknown. Therefore, while computing the combining vector to estimate users' signal, the correlation of compression noise elements at each AP is ignored ($\mathbf{Q}_{el}, \forall l$ is assumed to be a diagonal matrix), which will adversely affect the estimation quality. The sequential estimation of user signals results in equations (19) and (20) with $\mathbf{Z}_v$ in (8) replaced with the diagonal matrix $\mathbf{Z}_e^d = \text{blkdiag}(\mathbf{Z}_{e1}^d, \dots, \mathbf{Z}_{eL}^d) = \text{blkdiag}(\mathbf{Q}_{e1}^d + \sigma^2\mathbf{I}_N, \dots, \mathbf{Q}_{eL}^d + \sigma^2\mathbf{I}_N)$,

$$\hat{\mathbf{s}}_e = \mathbf{V}_e\hat{\mathbf{y}}_e, \tag{19}$$

where combining matrix $\mathbf{V}_e$ is formulated as follows:

$$\mathbf{V}_e = (\mathbf{H}^{\mathsf{H}}(\mathbf{Z}_e^d)^{-1}\mathbf{H} + \frac{1}{p}\mathbf{I}_K)^{-1}\mathbf{H}^{\mathsf{H}}(\mathbf{Z}_e^d)^{-1}. \tag{20}$$

The maximization of the sum-SE in option 2 follows the same steps as in Section III-A. The user's sum-SE using EC can be simplified as follows:

$$R_{EC} = \frac{\tau_u}{\tau_c} \log_2 \det(p\mathbf{H}\mathbf{H}^{\mathsf{H}}(\mathbf{Z}_e^d)^{-1} + \mathbf{I}_{NL})$$
$$\overset{(a)}{\le} \frac{\tau_u}{\tau_c} \sum_{l=1}^{L} \log_2 \det(p\mathbf{H}_l\mathbf{H}_l^{\mathsf{H}}(\mathbf{Z}_{el}^d)^{-1} + \mathbf{I}_N)$$
$$\overset{(b)}{\le} \frac{\tau_u}{\tau_c} \sum_{l=1}^{L} \log_2 \det(p\,\text{diag}(\mathbf{H}_l\mathbf{H}_l^{\mathsf{H}})(\mathbf{Z}_{el}^d)^{-1} + \mathbf{I}_N)$$
$$= \frac{\tau_u}{\tau_c} \sum_{l=1}^{L} \log_2 \det(p\mathbf{W}_l(\mathbf{Q}_{el}^d + \sigma^2\mathbf{I})^{-1} + \mathbf{I}_N), \tag{21}$$

where $\mathbf{W}_l$ is defined as $\mathbf{W}_l = \text{diag}(\|\mathbf{H}_{l[1,:]}\|^2, \dots, \|\mathbf{H}_{l[N,:]}\|^2)$. In (21), $\overset{(a)}{\le}$ and $\overset{(b)}{\le}$ are proved similar to $\overset{(a)}{\le}$ in (9).

The optimization problem to find the diagonal elements of $\mathbf{Q}_{el}^d$ and, subsequently, the number of bits to compress

each of the elements of the local received signal vector $\mathbf{y}_l$ is formulated as follows:

$$\arg \max_{(\mathbf{Q}_{el}^d)^{-1} \succeq 0} \quad \log_2 \det(p\mathbf{W}_l(\mathbf{Q}_{el}^d + \sigma^2 \mathbf{I}_N)^{-1} + \mathbf{I}_N)$$

$$\text{s.t.} \quad C_s = \log_2 \det((\mathbf{Q}_{el}^d)^{-1}\mathbf{P}_l + \mathbf{I}_N), \quad (22)$$

with $\mathbf{P}_l$ defined in (15). Note that $\mathbf{P}_l = p\mathbf{W}_l + \sigma^2\mathbf{I}_N$. Following a similar derivation as in Section III-A, the ith diagonal element of matrix $(\mathbf{Q}_{el}^d)^{-1}$, shown as $\frac{1}{\sigma_{eli}^2}$, can be calculated as follows:

$$\frac{1}{\sigma_{eli}^2} = \left[\frac{1}{\mu^*}\left(\frac{1}{\sigma^2} - \frac{1}{\mathbf{P}_{l[i,i]}}\right) - \frac{1}{\sigma^2}\right]^+. \quad (23)$$

IV. MEMORY CAPACITY MODEL

Regarding the memory capacity at the APs, two general scenarios are considered.

- **Fixed per AP (FAP)**: There is a fixed memory capacity C_{AP} per AP. Therefore, the total memory capacity depends on the number of APs.
- **Fixed total (FT)**: There is a fixed total memory capacity C_T that is divided among APs. Therefore, the memory capacity allocated to each AP depends on the number of APs.

We assume that received signal vectors are stored in on-chip cache memory [18]. Cache memory is desirable for its fast accessibility and low energy consumption. In reality, local CSI should also be stored in the memory. However, as the amount of data related to CSI is similar in each AP, we ignore the low precision storage of the local CSI. Note that the memory capacity available at one AP is shared among all the sub-carriers.

V. SIMULATION RESULTS

This section presents simulation results, which give insight into how the limited memory capacity in each AP can affect the average per-user SE. The simulation area is square with a perimeter of $D = 500$m. The APs are located on the perimeter of the area, and the distance between any two APs is the same. The users are uniformly located in a concentric square with a perimeter of 400m. The vertical distance between a user and an AP is 5m [4]. The total number of antennas is $NL = 128$ which are distributed in $L = \{2, 4, 8, 16, 32, 64, 128\}$ APs. The path-loss model of an urban microcell with 2GHz carrier frequency is considered [4], [24]. Accordingly, the large-scale fading coefficient is defined as follows:

$$\beta_{kl} = -30.5 - 36.7\log_{10}(\frac{d_{kl}}{1\text{m}}), \quad (24)$$

where d_{kl} and β_{kl} are the distance and large scale fading coefficient between user k and AP l, repectively. The noise variance at the APs is $\sigma^2 = -85$dBm, and the users' transmit power is $p = 10$mWatt.

In this section, $\frac{\mathbb{E}\{R_{VC}\}}{K}$ and $\frac{\mathbb{E}\{R_{EC}\}}{K}$ versus the number of APs are plotted. The expectations are with respect to all kinds of randomness. Scaling factor $\frac{\tau_u}{\tau_c}$ exists in both (9) and (21),

Fig. 2. Average per-user SE comparison using FAP memory model in a daisy chain fronthaul topology with two different numbers of users: (Left) $K = 4$. (Right) $K = 64$.

and as we don't consider any specific values for τ_u and τ_c, the factor is omitted while plotting the simulation results.

To make the simulation results clear, we start with an example. In the simulation figures, in case of $\{L = 64, N = 2\}$, the last AP stores $63F$ **two**-dimensional received signal vectors in the memory, and in case of $\{L = 128, N = 1\}$, the last AP stores $127F$ received signal **scalars** (as there is one antenna per AP), according to Section II-B. The total number of scalars to be stored in the last AP in the two cases mentioned are close to each other (i.e., $2 \times 63F$ versus $127F$). Therefore, it would seem that the compression noise should be almost the same. However, using VC and in the case of $\{L = 64, N = 2\}$, the last AP compresses scalars two by two ($63F$ pair of scalars), and in the case of $\{L = 128, N = 1\}$, the last AP compresses each scalar individually. Jointly compressing every two scalars in $\{L = 64, N = 2\}$ allows the AP to use the available memory more efficiently than $\{L = 128, N = 1\}$. Consequently, when the memory capacity is limited (e.g., $C_{AP} = 64KB$ in FAP), even though the number of scalars to be stored in the memory of the last AP in both cases is almost the same, $\{L = 64, N = 2\}$ outperforms $\{L = 128, N = 1\}$ even with the reduced macro-diversity. In other words, macro diversity in the case of $\{L = 128, N = 1\}$ can not compensate for the adverse effect of compression noise on average per-user SE. The above comparison can also be made between any other values of L. A similar conclusion can be drawn for EC.

In Fig. 2, it is observed that, under the assumption of infinite memory capacity, the distribution of the antennas in single-antenna APs improves the average per-user SE, especially when the number of users is large, e.g., $K = 64$. However, with the realistic assumption of limited memory capacity in each AP and using VC, the distribution of the antennas in single-antenna APs not only does not help in average per-user SE improvement but also reduces the average per-user SE, e.g., when $C_{AP} = 64KB$.

In Fig. 3, FT memory model is considered, and a similar trend to Fig. 2 is observed. Unlike FAP, in FT memory model, adding APs doesn't increase the total memory but makes the

Fig. 3. Average per-user SE Comparison using FT memory model in a daisy chain fronthaul topology with two different numbers of users: (Left) $K = 4$, (Right) $K = 64$.

memory per AP smaller. Furthermore, it is observed that the performance improvement of VC over EC is relatively small in this case, especially for the case of a low number of APs. This is because when the number of APs is low, each AP receives a large share of the total memory, reducing the difference between EC and VC.

VI. CONCLUSIONS

This paper discusses sequential processing with limited memory APs in a cell-free massive MIMO network with daisy chain fronthaul topology. The paper shows a trade-off between achieving macro diversity by distributing the antennas as much as possible in more APs and reducing the adverse effect of compression noise by distributing the antennas in fewer APs. Specifically, based on simulation results, distributing the antennas can benefit the average per-user SE when there is no memory limit at the APs, especially when the number of users is high. However, this is not the case when we limit the memory available at each AP to store the received signal vectors. With limited memory capacity constraints at each AP, the antennas tend to be collocated in fewer APs. Hence, the memory capacity highly impacts the number of APs among which the available antennas should be distributed.

VII. ACKNOWLEDGEMENT

REFERENCES

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

[2] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[3] Z. H. Shaik, E. Björnson, and E. G. Larsson, "Cell-Free Massive MIMO with Radio Stripes and Sequential Uplink Processing," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, 2020.

[4] Z. H. Shaik, E. Björnson, and E. G. Larsson, "MMSE-Optimal Sequential Processing for Cell-Free Massive MIMO With Radio Stripes," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7775–7789, 2021.

[5] K. W. Helmersson, P. Frenger, and A. Helmersson, "Uplink D-MIMO Processing Using Kalman Filter Combining," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pp. 1703–1708, 2022.

[6] V. Ranjbar, S. Pollin, and M. Moonen, " Finite Precision Implementation of Recursive Algorithms for Uplink Detection in Cell-Free Networks," in *2022 IEEE Globecom Workshops (GC Wkshps)*, pp. 25–30, 2022.

[7] E. Björnson and L. Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247–4261, 2020.

[8] E. Björnson and L. Sanguinetti, "Making Cell-Free Massive MIMO Competitive With MMSE Processing and Centralized Implementation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 77–90, 2020.

[9] E. Björnson, L. Sanguinetti, and J. Hoydis, "Hardware Distortion Correlation Has Negligible Impact on UL Massive MIMO Spectral Efficiency," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1085–1098, 2019.

[10] X. Hu, C. Zhong, X. Chen, W. Xu, H. Lin, and Z. Zhang, "Cell-Free Massive MIMO Systems With Low Resolution ADCs," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 6844–6857, 2019.

[11] D. Verenzuela, E. Björnson, and M. Matthaiou, "Optimal Per-Antenna ADC Bit Allocation in Correlated and Cell-Free Massive MIMO," *IEEE Transactions on Communications*, vol. 69, no. 7, pp. 4767–4780, 2021.

[12] Y. Xiong, S. Sun, N. Wei, L. Liu, and Z. Zhang, "Asymptotic Analysis for Cell-Free Massive MIMO With MMSE Combining and Low-Resolution ADCs," *IEEE Communications Letters*, vol. 25, no. 10, pp. 3219–3223, 2021.

[13] Y. Xiong, S. Sun, L. Liu, Z. Zhang, and N. Wei., "Performance Analysis and Bit Allocation of Cell-Free Massive MIMO Network With Variable-Resolution ADCs," *IEEE Transactions on Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.

[14] M. Bashar, H. Q. Ngo, K. Cumanan, A. G. Burr, P. Xiao, E. Björnson, and E. G. Larsson, "Uplink Spectral and Energy Efficiency of Cell-Free Massive MIMO With Optimal Uniform Quantization," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 223–245, 2021.

[15] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao, "Energy Efficiency of the Cell-Free Massive MIMO Uplink With Optimal Uniform Quantization," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 4, pp. 971–987, 2019.

[16] H. Masoumi and M. J. Emadi, "Performance Analysis of Cell-Free Massive MIMO System With Limited Fronthaul Capacity and Hardware Impairments," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1038–1053, 2020.

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

[18] J. Rodríguez Sánchez, *Systems with Massive Number of Antennas: Distributed Approaches*. PhD thesis, Electrical and Information Technology, Lund University, 2022.

[19] T. report, "Arm Neoverse E1 Core, Technical Reference Manual," 2019. https://developer.arm.com/documentation/101560/latest.

[20] T. M. Cover and J. A. Thomas, *Elements of information theory*. Hoboken, NJ, USA: Wiley, 2012.

[21] 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 Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1555–1567, 2014.

[22] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2 ed., 2012.

[23] A. del Coso and S. Simoens, "Distributed compression for MIMO coordinated networks with a backhaul constraint," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4698–4709, 2009.

[24] 3GPP, "Further advancements for E-UTRA physical layer aspects (Release 9)," *3GPP TS 36.814*, Mar. 2017.