Sequential Processing Strategies in Fronthaul Constrained Cell-Free Massive MIMO Networks

Vida Ranjbar, *Member, IEEE*, Robbert Beerten, *Graduate Student Member, IEEE*, Marc Moonen, *Fellow, IEEE*, and Sofie Pollin, *Senior Member, IEEE*

***Abstract*—In a cell-free massive MIMO (CFmMIMO) network with a daisy-chain fronthaul, the amount of information that each access point (AP) needs to communicate with the next AP in the chain is determined by the location of the AP in the sequential fronthaul. Therefore, we propose two sequential processing strategies to combat the adverse effect of fronthaul compression on the sum of users' spectral efficiency (SE): 1) linearly increasing fronthaul capacity allocation among APs and 2) Two-Path users' signal estimation. The two strategies show superior performance in terms of sum SE compared to the equal fronthaul capacity allocation and Single-Path sequential signal estimation.**

***Index Terms*—Fronthaul compression in cell-free massive MIMO networks, sequential fronthaul topology.**

I. Introduction

SEQUENTIAL processing in CFmMIMO networks has attracted significant research attention in recent years, as evidenced by numerous studies such as [1], [2], [3], [4], [5] among others. In a sequential network, the APs take turns in refining users' signal estimate in the uplink direction. The architecture reduces the total fronthaul signaling among the APs and the CPU and has a low practical implementation complexity [2], which makes it the favored architecture in some particular scenarios such as railways, factories, museums, etc. Using Kalman filtering, it can be proved that the performance of a distributed sequential uplink signal estimation at the APs is the same as the performance of the centralized linear minimum mean square error (LMMSE) signal estimation [2], [6], under ideal condition such as no fronthaul compression requirement.

Fronthaul signaling among APs/CPU over limited-capacity (wired or wireless) links constitutes an integral part of the whole CFmMIMO networking. The performance degradation due to the limited-capacity fronthaul links can be mitigated in several ways [3], [4], [7], [8], among others. In a sequential architecture, the fact that APs can exchange information among each other can be exploited to indirectly lower the bit length of the exchanged users' signal estimates [9]. Different strategies, such as compare and forward (CNF) [3], [4], sequential mean squared error (MSE) minimization-based user-AP association [8] and sequential maximization of the local mutual information between the users' signals and information available at each AP under limited-capacity fronthaul links [10], are considered to mitigate the adverse effect of a limited-capacity fronthaul link on users' SE. The authors in [11], considered the downlink CFmMIMO with serial fronthaul and took an information-theoretic approach towards fronthaul capacity allocation among APs. They demonstrated that having a total fronthaul capacity, "water-pouring" distribution of the total fronthaul capacity over fronthaul links among APs will optimize the channel capacity.

Motivation: In a sequential CF-mMIMO architecture, the amount of information an AP sends to the next AP depends on its position in the fronthaul chain. AP l refines the users' signal estimates received from AP $l-1$ (i.e., $\tilde{\mathbf{s}}_{l-1}$) using its local received vector $\mathbf{y}_l$, and then forwards the compressed refined users' signal estimate $\tilde{\mathbf{s}}_l$ to AP $l+1$. The estimate thus passes through APs $1,\ldots,l$ before it reaches AP $l+1$. Consequently, each AP must convey the information embedded in all preceding APs' received signals to the next AP/CPU. As the index l increases, the forwarded estimate $\tilde{\mathbf{s}}_l$ reflects processing over an increasing number of antennas. Therefore, allocating equal fronthaul capacity to all APs is inefficient, since early APs may not use their full capacity while later APs require more to forward their refined estimates with lower compression noise.

Contributions: In this letter, we study the uplink sum SE performance of a practical limited-capacity sequential fronthaul CFmMIMO network. We propose two fronthaul deployment strategies that can significantly improve the sum SE. The first strategy considers non-uniform linearly increasing fronthaul allocation among APs under the assumption of a fixed total number of bits per uplink sample for fronthaul signaling. One application scenario for linearly increasing fronthaul capacity is when the fronthaul links between APs are wireless, and there is a particular frequency bandwidth that is allocated for wireless fronthaul signaling. The amount of spectrum that is allocated to each AP for fronthaul signaling increases with the AP index. Finally, a combination of Two-path signal processing with linearly increasing fronthaul gives the highest sum SE, especially for a higher number of users. MATLAB source code can be found in https://gitlab.com/VidaRanjbar/Sequential-fronthaul-constrained-CFmMIMO_WCL2026.

Notations: We denote vectors and matrices with boldface lower-case and upper-case letters, respectively. Transpose and

Received 8 December 2025; accepted 23 January 2026. Date of publication 3 February 2026; date of current version 16 February 2026. This work was supported by the European Union's Horizon 2020 Research and Innovation Program through 6G-BRICKS under Grant 101096954, and in part by the Research Foundation—Flanders (FWO) under Grant G0C0623N. The associate editor coordinating the review of this article and approving it for publication was S. Dang. *(Corresponding author: Vida Ranjbar.)*

Vida Ranjbar is with the Institute of Networked and Embedded Systems, University of Klagenfurt, 9020 Klagenfurt, Austria (e-mail: vida.ranjbar8@gmail.com).

Robbert Beerten, Marc Moonen, and Sofie Pollin are with the Department of Electrical Engineering, KU Leuven, 3000 Leuven, Belgium (e-mail: robbert.beerten@kuleuven.be; marc.moonen@kuleuven.be; sofie.pollin@kuleuven.be).

Digital Object Identifier 10.1109/LWC.2026.3661438

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

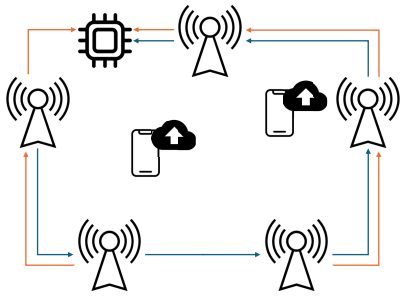

Fig. 1. Sequential uplink signal estimation. User signal estimates are refined following **either** the blue path until it reaches CPU (Single-Path (SP) signal estimation), **or** it is refined following two short orange paths, and the two resulting signal estimates from the two paths get fused in the CPU (Two-Path (TP) signal estimation).

conjugate transpose operations are denoted by ^{T} and ^{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 $\mathbf{x}$. *diag*($\mathbf{x}$) is a diagonal matrix with the same diagonal elements as the elements of vector $\mathbf{x}$. $\mathbf{X} \succeq \mathbf{0}$ means $\mathbf{X}$ is positive semi-definite.

II. System Model

This letter considers the uplink of a CFmMIMO network with APs connected in a sequential fronthaul, as shown in Fig. 1. There are K single-antenna users in the network served by L APs, each equipped with N antennas. The total number of antennas is $M = LN$. AP l refines the compressed users' signal estimates received from AP $l-1$, i.e., $\tilde{\mathbf{s}}_{l-1}$, based on its own received signal vector in the uplink, formulated as:

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

where $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}, p\mathbf{I}_K)$ is users' signal vector, $\mathbf{n}_l \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_N)$ is the noise vector at AP l. Each column of matrix $\mathbf{H}_l$ follows a circularly symmetric complex Gaussian distribution, i.e., $\mathbf{H}_{l[:,k]} \sim \mathcal{CN}(\mathbf{0}, \mathbf{R}_{kl})$. We don't elaborate on channel estimation and refer to [12] for the channel estimation procedure. It is assumed that the local channel realizations are known at the APs. The local channel realizations remain constant during one coherence block of time and frequency, with a total τ_c sample per block. The phase and carrier frequency synchronization between distributed APs to enable coherent transmission and reception in time-division duplex (TDD) CFmMIMO is discussed in [13]. In this letter, perfect synchronization is assumed among APs.

The two mentioned sources of information, i.e., the users' signal estimate received from AP $l-1$ and the local received signal vector at AP l, are combined based on the local channel matrix, i.e., $\mathbf{H}_l$, and the error statistics of the estimates received from AP $l-1$, i.e., $\mathbf{C}_{l-1} = \mathbb{E}\{\mathbf{e}_{l-1}\mathbf{e}_{l-1}^{\mathrm{H}}\}$, $\mathbf{e}_{l-1} = \mathbf{s} - \tilde{\mathbf{s}}_{l-1}$. $\tilde{\mathbf{s}}_{l-1}$ is formulated as follows:

$$\tilde{\mathbf{s}}_{l-1} = \hat{\mathbf{s}}_{l-1} + \mathbf{q}_{l-1}, \tag{2}$$

where $\hat{\mathbf{s}}_{l-1}$ is the refined users' signal estimate at AP $l-1$ before compression, and $\mathbf{q}_{l-1}$ is the compression noise vector at AP $l-1$, independent from $\hat{\mathbf{s}}_{l-1}$ and imposed by the limited-capacity fronthaul link that connects AP $l-1$ to AP l. Based on standard results of information theory in [14, Ch. 3], $\mathbf{q}_l \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_l)$. Using Kalman filtering concept [2], [6], the refined signal estimate at AP l is as follows:

$$\begin{aligned}\hat{\mathbf{s}}_l &= \tilde{\mathbf{s}}_{l-1} + \mathbf{\Gamma}_l(\mathbf{y}_l - \mathbf{H}_l\tilde{\mathbf{s}}_{l-1}) \\ &= \tilde{\mathbf{s}}_{l-1} + \mathbf{\Gamma}_l(\mathbf{H}_l\mathbf{e}_{l-1} + \mathbf{n}_l),\end{aligned} \tag{3}$$

where $\mathbf{\Gamma}_l$ is the $K \times N$ combining matrix computed as follows:

$$\mathbf{\Gamma}_l = \mathbf{C}_{l-1}\mathbf{H}_l^{\mathrm{H}}\left(\mathbf{H}_l\mathbf{C}_{l-1}\mathbf{H}_l^{\mathrm{H}} + \sigma^2\mathbf{I}_N\right)^{-1}. \tag{4}$$

At AP 1, $\mathbf{C}_0 = p\mathbf{I}_K$ and $\tilde{\mathbf{s}}_0 = \mathbf{0}_{K\times 1}$. Based on (2) and (3), $\tilde{\mathbf{s}}_l$ can be formulated as follows:

$$\begin{aligned}\tilde{\mathbf{s}}_l &= \hat{\mathbf{s}}_l + \mathbf{q}_l \\ &= \begin{bmatrix}\mathbf{I}_K - \mathbf{\Gamma}_l\mathbf{H}_l & \mathbf{\Gamma}_l\end{bmatrix}\begin{bmatrix}\hat{\mathbf{s}}_{l-1} \\ \mathbf{y}_l\end{bmatrix} + (\mathbf{I}_K - \mathbf{\Gamma}_l\mathbf{H}_l)\mathbf{q}_{l-1} + \mathbf{q}_l.\end{aligned} \tag{5}$$

By recursion, $\tilde{\mathbf{s}}_l$ at AP l can be rewritten as the combination of received signal vectors of AP 1 to AP l and an additive collective compression noise vector that was added to the users' signal estimates through the sequence until AP l:

$$\begin{aligned}\tilde{\mathbf{s}}_l &= \begin{bmatrix}\mathbf{V}_{1l} & \cdots & \mathbf{V}_{ll}\end{bmatrix}\begin{bmatrix}\mathbf{y}_1^T & \cdots & \mathbf{y}_l^T\end{bmatrix}^T \\ &\quad + \underbrace{\begin{bmatrix}\mathbf{A}_{1l} & \cdots & \mathbf{A}_{ll}\end{bmatrix}\begin{bmatrix}\mathbf{q}_1^T & \cdots & \mathbf{q}_l^T\end{bmatrix}^T}_{\mathbf{q}_l^{\mathrm{eff}}} \\ &= \sum_{i=1}^{l}\mathbf{V}_{il}\mathbf{H}_i\mathbf{s} + \mathbf{V}_{il}\mathbf{n}_i + \mathbf{A}_{il}\mathbf{q}_i,\end{aligned} \tag{6}$$

where

$$\mathbf{V}_{il} = \begin{cases}\left(\prod_{j=i+1}^{l}(\mathbf{I}_K - \mathbf{\Gamma}_j\mathbf{H}_j)\right)\mathbf{\Gamma}_i, & \text{for } i < l \\ \mathbf{\Gamma}_l, & \text{for } i = l\end{cases} \tag{7}$$

and

$$\mathbf{A}_{il} = \begin{cases}\prod_{j=i+1}^{l}(\mathbf{I}_K - \mathbf{\Gamma}_j\mathbf{H}_j), & \text{for } i < l \\ \mathbf{I}_K, & \text{for } i = l\end{cases} \tag{8}$$

where the order of matrix multiplications is AP l, AP $l-1$, …, AP $i+1$ from left to right. It is easily shown that $\mathbf{V}_{il} = \{(\mathbf{I}_K - \mathbf{\Gamma}_l\mathbf{H}_l)\mathbf{V}_{i(l-1)}, \forall i \le l-1\}$ and $\mathbf{A}_{il} = \{(\mathbf{I}_K - \mathbf{\Gamma}_l\mathbf{H}_l)\mathbf{A}_{i(l-1)}, \forall i \le l-1\}$. Finally, the error covariance matrix on $\tilde{\mathbf{s}}_l$ is formulated as follows:

$$\begin{aligned}\mathbf{C}_l &= \mathbb{E}\{\mathbf{e}_l\mathbf{e}_l^{\mathrm{H}}\} \\ &= \mathbb{E}\{(\mathbf{s} - \tilde{\mathbf{s}}_l)(\mathbf{s} - \tilde{\mathbf{s}}_l)^{\mathrm{H}}\} \\ &= (\mathbf{I}_K - \mathbf{\Gamma}_l\mathbf{H}_l)\mathbf{C}_{l-1} + \mathbf{Q}_l,\end{aligned} \tag{9}$$

which can be proved considering (3), (4) and (5). Along with the uplink samples, the covariance matrices $\mathbf{C}_{l-1}$ should also be exchanged with AP l. However, as this matrix will remain constant over one coherence block and changes infrequently, we assume it is exchanged with high precision. If the sequential processing of the signal of User k is terminated at AP l and the rest of the APs only relay the estimate toward the CPU, an achievable SE for user k can be computed as follows:

$$\mathrm{SE}_{kl} = \frac{\tau_u}{\tau_c}\mathbb{E}\{\log_2(1 + \mathrm{SINR}_{kl})\}. \tag{10}$$

where SINR_{kl} is defined in (11), shown at the bottom of the next page. τ_c and τ_u are the number of samples per coherence block and number of samples for uplink data transmission per coherence block, respectively.

III. Fronthaul Compression Noise

As already mentioned, for a fixed fronthaul capacity at a particular AP, the user's signal estimates are refined,

compressed, and then sent to the next AP. In this letter, we use equal inter-user (EIU) fronthaul compression where a particular AP allocates the same number of bits to each user's signal. This is an element-wise compression scheme. Alternatively, we consider fronthaul compression based on sum compression noise variance minimization (SCNM), formulated as minimization of the trace of the compression noise covariance matrix in [4]. Lastly, we introduce a weighted sum interference plus noise variance minimization-based fronthaul compression method (WSINM). The last two compression methods are categorized as vector-wise compression.

Based on (2), (3), (4), and (9), the correlation matrix of $\hat{\mathbf{s}}_l$ (refined signal before compression at AP l) is defined as:

$$\begin{aligned}\mathbf{P}_l &= \mathbb{E}\{\hat{\mathbf{s}}_l\hat{\mathbf{s}}_l^{\mathrm{H}}|\mathbf{C}_{l-1},\mathbf{H}_l\}\\ &= \mathbf{P}_{l-1}+\mathbf{Q}_{l-1}+\mathbf{\Gamma}_l\mathbf{H}_l\mathbf{C}_{l-1}-\mathbf{Q}_{l-1}\mathbf{H}_l^{\mathrm{H}}\mathbf{\Gamma}_l^{\mathrm{H}}-\mathbf{\Gamma}_l\mathbf{H}_l\mathbf{Q}_{l-1}. \quad (12)\end{aligned}$$

The derivations are not detailed due to space limitations. However, it is easily verifiable. Note that $\mathbb{E}\{\tilde{\mathbf{s}}_{l-1}\mathbf{e}_{l-1}^{\mathrm{H}}\} = -\mathbf{Q}_{l-1}$.

A. Element-Wise Compression: Equal Inter-User Fronthaul Capacity Allocation

If there are b_{kl} bits to compress the kth complex valued element of $\hat{\mathbf{s}}_l$ at AP l, based on standard results of rate distortion theory [14, Ch. 3], the relation between b_{kl} and the kth diagonal element of matrices $\mathbf{P}_l$ and $\mathbf{Q}_l$ is as follows:

$$b_{kl} = \log_2\Big(\mathbf{P}_{l[k,k]}\mathbf{Q}_{l[k,k]}^{-1}+1\Big) \rightarrow \mathbf{Q}_{l[k,k]} = \frac{\mathbf{P}_{l[k,k]}}{2^{b_{kl}}-1}. \quad (13)$$

In the simulation section, we consider $b_{kl} = R_l/K$ for EIU, with R_l being the number of bits/sample allocated to AP l for fronthaul signaling of $\hat{\mathbf{s}}_l$.

B. Vector-Wise Compression: Sum Compression Noise Variance Minimization

To compute the compression noise covariance matrix $\mathbf{Q}_l$ at a particular AP l, a sum compression noise variance minimization (SCNM) problem in [4] is formulated as follows:

$$\begin{aligned}\arg\min_{\mathbf{Q}_l\succeq\mathbf{0}} \quad & \sum_k \mathbf{Q}_{l[k,k]}\\ \text{s.t.} \quad & R_l = \log_2\det\Big(\mathbf{P}_l\mathbf{Q}_l^{-1}+\mathbf{I}_K\Big). \quad (14)\end{aligned}$$

Problem (14) can be solved using [4, Proposition 1].

C. Vector-Wise Compression: Weighted Sum Interference Plus Noise Variance Minimization

Alternatively, inspired by the weighted MMSE (WMMSE) problem in [15], a weighted sum interference plus noise variance minimization (WSINM) problem is formulated to compute the compression noise covariance matrix, as follows:

$$\begin{aligned}\arg\min_{\mathbf{Q}_l\succeq\mathbf{0},w_{lk}\forall k} \quad & \sum_k w_{lk}X_{lk} - \sum_k \log_2 w_{lk}\\ \text{s.t.} \quad & R_l = \log_2\det\Big(\mathbf{P}_l\mathbf{Q}_l^{-1}+\mathbf{I}_K\Big). \quad (15)\end{aligned}$$

where X_{kl} is defined in (11). The above problem is not jointly convex with respect to $\mathbf{Q}_l$ and $\{w_{lk},\forall k\}$. Hence, the block coordinate descent (BCD) procedure is used to iteratively optimize the multi-variable function, as follows:

- For a fixed $\{w_{lk},\forall k\}$, the problem (15) reduces to the minimization of a weighted trace of the compression noise covariance matrix that can be solved similarly to the minimization strategy in problem (14). More specifically, for a fixed $\{w_{lk},\forall k\}$, matrix $\mathbf{W}_l$ is defined as:
$$\mathbf{W}_l = diag(w_{l1},\ldots,w_{lK}), \quad (16)$$
and the problem in (15) is equivalent to:
$$\begin{aligned}\arg\min_{\mathbf{Q}_l\succeq\mathbf{0}} \sum_{k=1}^{K} w_{lk}\mathbf{Q}_{l[k,k]} \quad &= tr(\mathbf{W}_l\mathbf{Q}_l)\\ &= tr\Bigg(\underbrace{\mathbf{W}_l^{\frac{1}{2}}\mathbf{Q}_l\mathbf{W}_l^{\frac{\mathrm{H}}{2}}}_{\bar{\mathbf{Q}}_l}\Bigg)\\ \text{s.t.} \quad R_l = \log_2\det\Bigg(&\underbrace{\mathbf{W}_l^{\frac{1}{2}}\mathbf{P}_l\mathbf{W}_l^{\frac{\mathrm{H}}{2}}}_{\bar{\mathbf{P}}_l}\underbrace{\mathbf{W}_l^{\frac{-1}{2}}\mathbf{Q}_l^{-1}\mathbf{W}_l^{\frac{-\mathrm{H}}{2}}}_{\bar{\mathbf{Q}}_l^{-1}}+\mathbf{I}_K\Bigg), \quad (17)\end{aligned}$$
where the properties of the trace, determinant, and the square root of a matrix are used [16].
- For a fixed $\mathbf{Q}_l$ the problem in (15) is concave with respect to $\{w_{lk},\forall k\}$ and the optimal $\{w_{lk},\forall k\}$ are calculated as:
$$w_{lk}^* = \frac{1}{\ln 2\times X_{lk}},\forall k. \quad (18)$$

The iteration continues until convergence. The convergence proof of the iterative optimization is straightforward using the fact that the objective function is decreasing in each iteration and it is bounded from below for $\{X_{lk}\neq 0,\forall k\}$, which is almost always the case in practice. In other words, in practice, the interference plus noise is always non-zero for all users in a multi-user CFmMIMO network.

If $\{w_{lk}^*,\forall k\}$ are inserted back into the problem (15), we have an equivalent problem as follows:

$$\begin{aligned}\arg\max_{\mathbf{Q}_l\succeq\mathbf{0}} \quad & \sum_k \log_2\frac{1}{X_{kl}}\\ \text{s.t.} \quad & R_l = \log_2\det\Big(\mathbf{P}_l\mathbf{Q}_l^{-1}+\mathbf{I}_K\Big). \quad (19)\end{aligned}$$

$$\mathrm{SINR}_{kl} = \frac{p|\sum_{i=1}^{l}\mathbf{V}_{il[k,:]}\mathbf{H}_{i[:,k]}|^2}{\underbrace{\sum_{j\neq k}p|\sum_{i=1}^{l}\mathbf{V}_{il[k,:]}\mathbf{H}_{i[:,j]}|^2+\sigma^2\sum_{i=1}^{l}\mathbf{V}_{il[k,:]}\mathbf{V}_{il[k,:]}^{\mathrm{H}}+\sum_{i=1}^{l-1}\mathbf{A}_{il[k,:]}\mathbf{Q}_i\mathbf{A}_{il[k,:]}^{\mathrm{H}}+\mathbf{Q}_{l[k,k]}}}_{X_{kl}}. \quad (11)$$

Based on (10) and (11), the objective function in (19) differs from sum SE up to a constant (the numerator in (11)) in the high SINR regime. Hence, (19) is equivalent to the sum SE maximization problem in that regime.

IV. Fronthaul Capacity Allocation

In an uplink sequential signal estimation without any constraint on the capacity of the fronthaul links and the need for compression, the number of bits forwarded from one AP to the next increases with AP index. In the presence of limited-capacity fronthaul, the APs in the beginning need less capacity than the APs toward the end of the sequence. If there is a total fronthaul capacity (for example, there is a fixed number of subcarriers allocated to wireless fronthaul signaling), then the capacity allocation to each AP should take into account the increasing need for fronthaul signaling capacity toward the end of the sequence. This means that equal capacity allocation to the APs might be suboptimal, especially when the number of users is large, as the first AP may not need the capacity allocated to it, while the last AP may need to compress the users' signals estimate heavily. We compare two fronthaul capacity allocations among APs:

- Equal fronthaul allocation (EF): If the total number of bits per uplink sample of users' signal is R_T, the bit allocated to each AP is $R_l = \frac{R_T}{L}$.
- Linearly increasing fronthaul capacity allocation (LF): In this approach, the fronthaul capacity increases linearly with AP index, $R_l = \frac{2R_T}{L(L+1)}l$.

The fronthaul capacity allocation scheme can also be logarithmic, i.e., $R_l = \frac{\log_2 l}{\sum_i \log_2 i} R_T$. This capacity allocation strategy typically exhibits inferior performance compared to the linearly increasing fronthaul capacity allocation in the considered scenario and is therefore removed from the simulation section.

V. Two-Path Sequential Signal Propagation

Based on (5), the signal estimate in AP l has the combined effect of compression noise from AP 1 to AP l as well as the corresponding received signal vectors. Hence, the longer the sequence, the bigger the combined effect of compression noise on the users' signal estimate as well as the amount of information to be exchanged with the next AP.

To compensate for this combined effect of compression noise, the linear increase in the fronthaul capacity along the sequence is proposed in Section IV. An alternative to increasing the fronthaul capacity proposal is the Two-Path signal propagation, as shown in Fig. 1.

The effect of increasing compression noise when fronthaul capacity is not growing sufficiently can be mitigated if the users' signal estimate propagates through two (instead of one) paths to the CPU, and then in the CPU, the local estimates from the two paths can be combined to estimate users' signals globally. In this case, the combined compression noise effect in each AP is at most from $L/2$ preceding APs instead of L APs (with the assumption of equally long paths). From Path $r \in \{1, 2\}$, the CPU receives the following signal estimate:

$$\tilde{\mathbf{s}}^r = \begin{bmatrix}\mathbf{V}_1^r & \cdots & \mathbf{V}_{L_r}^r\end{bmatrix}\begin{bmatrix}\mathbf{y}_1^{rT} & \cdots & \mathbf{y}_{L_r}^{rT}\end{bmatrix}^T + \underbrace{\begin{bmatrix}\mathbf{A}_1^r & \cdots & \mathbf{A}_{L_r}^r\end{bmatrix}\begin{bmatrix}\mathbf{q}_1^{rT} & \cdots & \mathbf{q}_{L_r}^{rT}\end{bmatrix}^T}_{\mathbf{q}^{r-\text{eff}}}, \tag{20}$$

where L_r is the number of APs involved in sequential processing in Path r and $L = \sum_{r=1}^{2} L_r$. The effective channel and effective noise from path r is defined as $\mathbf{G}^r = \sum_{l=1}^{L_r} \mathbf{V}_l^r \mathbf{H}_l^r$ and $\mathbf{z}^r = \sum_{l=1}^{L_r} \mathbf{V}_l^r \mathbf{n}_l^r + \mathbf{A}_l^r \mathbf{q}_l^r \sim \mathcal{CN}(\mathbf{0}, \mathbf{Z}^r)$. Covariance matrix $\mathbf{Z}^r$ is formulated as:

$$\mathbf{Z}^r = \sum_{l=1}^{L_r} \sigma^2 \mathbf{V}_l^r \mathbf{V}_l^{r\mathrm{H}} + \mathbf{A}_l^r \mathbf{Q}_l^r \mathbf{A}_l^{r\mathrm{H}}. \tag{21}$$

The superscript r is meant to distinguish APs belonging to different paths. The overall information vector $\mathbf{y}^{\mathrm{TP}}$ (TP stands for Two-Path), effective channel matrix $\mathbf{G}^{\mathrm{TP}}$ and noise vector $\mathbf{z}^{\mathrm{TP}}$ from two paths are formulated as follows respectively:

$$\mathbf{y}^{\mathrm{TP}} = \begin{bmatrix}\tilde{\mathbf{s}}^1 \\ \tilde{\mathbf{s}}^2\end{bmatrix}, \quad \mathbf{G}^{\mathrm{TP}} = \begin{bmatrix}\mathbf{G}^1 \\ \mathbf{G}^2\end{bmatrix}, \quad \mathbf{z}^{\mathrm{TP}} = \begin{bmatrix}\mathbf{z}^1 \\ \mathbf{z}^2\end{bmatrix} \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{Z}^{\mathrm{TP}}\right),$$

where $\mathbf{Z}^{\mathrm{TP}} = blkdiag(\mathbf{Z}^1, \mathbf{Z}^2)$. The global LMMSE estimate of users' signal based on $\mathbf{y}^{\mathrm{TP}}$ in the CPU is as follows:

$$\hat{\mathbf{s}}^{\mathrm{TP}} = \mathbf{V}^{\mathrm{TP}} \mathbf{y}^{\mathrm{TP}}, \quad \mathbf{V}^{\mathrm{TP}} = p\mathbf{G}^{\mathrm{TP}\mathrm{H}} \left(p\mathbf{G}^{\mathrm{TP}} \mathbf{G}^{\mathrm{TP}\mathrm{H}} + \mathbf{Z}^{\mathrm{TP}}\right)^{-1}. \tag{22}$$

VI. Simulation Results

We validate the proposed approaches for sequential signal estimation in a daisy chain CFmMIMO network. A total number of $M = 120$ antennas are distributed among L APs. The APs are located on the perimeter of a circle with a radius of $300m$ meters. The users are uniformly distributed in an inner circle with a radius of $150m$. The number of users and APs is specified under each simulation figure. The large-scale fading coefficient between user k and AP l is as follows:

$$\beta_{kl} = -30.5 - 36.7 \log_{10}(d_{kl}), \tag{23}$$

following the propagation model of 3GPP Urban Microcell [17] with 2GHz carrier frequency and bandwidth of $B = 100$MHz [17]. Parameter d_{kl} is the distance between AP l and user k in meters. The noise variance is assumed to be $\sigma^2 = -85\mathrm{dBm}$. The users' transmit power is set to $p = 20\mathrm{dBm}$ (100mWatts). It is assumed that $\tau_u = \tau_c - \tau_p$, where $\tau_c = 200$ and $\tau_p = K$ samples are used in each coherence block to acquire channel knowledge. Suppose $R_T = 1000\mathrm{bit/sample}$, it means that for one uplink transmission (one uplink sample compression), there are 1000 bits to be shared between APs. If there are 4 APs, for example, in total, the linearly increasing fronthaul will allocate $R_1 = 100$, $R_2 = 200$, $R_3 = 300$ and $R_4 = 400$ bits to the APs 1 to 4. Afterwards, if the method in Section III-A is used for inter-user fronthaul capacity allocation, then AP 1 will allocate $b_{k1} = 100/K$ bits to each user, AP 2 will allocate $b_{k2} = 200/K$ bits to each user, etc. R_l can be converted to the equivalent bandwidth of the link connecting AP l to AP $l+1$=$R_l \times \#$ uplink sample/second.

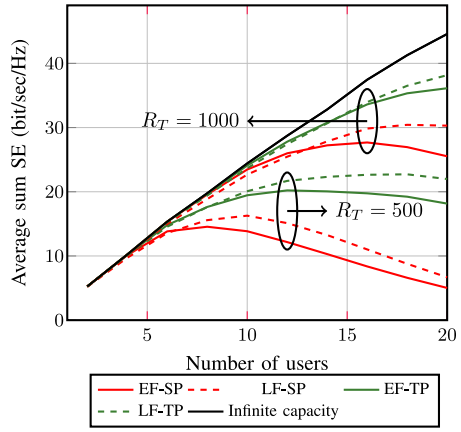

Fig. 2. Impact of limited capacity fronthaul link: sum SE vs number of users. For bit allocation among users, WSINM is considered. $L = 12$, $M = 120$.

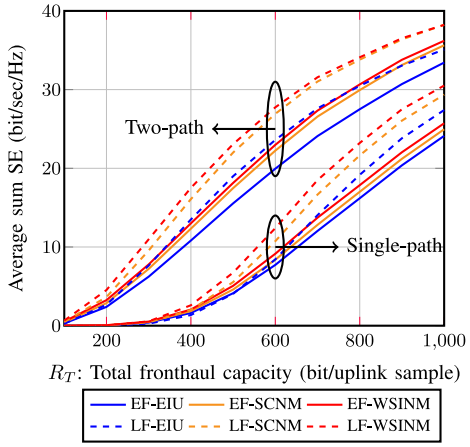

Fig. 3. Impact of limited capacity fronthaul link: sum SE vs total fronthaul capacity R_T. $L = 12$, $M = 120$ and $K = 20$.

Fig. 2 shows that when $K = 20$ and $R_T = 500$ bits/sample, linearly increasing fronthaul Two-Path signal estimation (LF-TP) and Equal fronthaul Two-Path signal estimation (EF-TP) will improve the sum SE performance by about 361% and 262% compared to the Equal Fronthaul Single-Path (EF-SP), respectively.

Having two paths for sequential signal processing and fusion in the CPU can mitigate the adverse effect of fronthaul compression on the sum users' SE, even with the most rudimentary fronthaul capacity allocation among APs (i.e., EF) and among users (i.e., EIU), as shown in Fig. 3.

VII. Conclusion

This letter shows that given a total fronthaul capacity in the cell-free massive MIMO (CFmMIMO) networks with sequential fronthaul, e.g., a total number of subcarriers allocated to perform wireless fronthaul signaling, the APs toward the end of the sequence should get a larger share of this total fronthaul capacity than the APs at the beginning of the sequence. The heterogeneous fronthaul capacity allocation enhances user spectral efficiency (SE), especially when the number of users is high. Ultimately, breaking the Single-Path signal estimation into Two-Path signal estimation can alleviate the detrimental effect of fronthaul compression, even when the equal fronthaul allocation among APs (EF) is considered. It is worth noting that the proposed strategies can be extended to other fronthaul topologies, as many fronthaul topologies can be reduced to a virtual sequential topology with virtual nodes (APs or clusters of APs) connected sequentially to each other. Future work can look into optimizing the number of APs per path in Two-Path signal propagation, and optimizing the fronthaul capacity allocation among APs instead of using heuristic approaches.

Acknowledgment

The resources and services used in this letter were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government.

References

[1] 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)*, 2020, pp. 1–6.

[2] 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.

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

[4] I. Chiotis and A. L. Moustakas, "Uplink performance optimization of limited-capacity radio stripes," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 12382–12395, Sep. 2024.

[5] S. Jo, H. Lee, and S.-H. Park, "Joint precoding and fronthaul compression for cell-free MIMO downlink with radio stripes," in *Proc. IEEE Glob. Commun. Conf.*, 2023, pp. 4945–4951.

[6] K. W. Helmersson, P. Frenger, and A. Helmersson, "Uplink D-MIMO processing using Kalman filter combining," in *Proc. IEEE Glob. Commun. Conf.*, 2022, pp. 1703–1708.

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

[8] I. Chiotis and A. L. Moustakas, "A sequential UE selection scheme for user-centric communications employing radio stripes," in *Proc. IEEE 25th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2024, pp. 161–165.

[9] V. Ranjbar, R. Beerten, M. Moonen, and S. Pollin, "Joint sequential fronthaul quantization and hardware complexity reduction in uplink cell-free massive MIMO networks," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, 2024, pp. 1–6.

[10] I. Chiotis and A. L. Moustakas, "Mutual information optimization for cell-free massive MIMO with sequential limited-rate fronthaul," in *Proc. IEEE Middle East Conf. Commun. Netw. (MECOM)*, 2024, pp. 145–150.

[11] 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, 2020.

[12] 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. [Online]. Available: https://www.emerald.com/ftsig/article-abstract/11/3-4/154/1328599/Massive-MIMO-Networks-Spectral-Energy-and-Hardware?redirectedFrom=fulltext

[13] U. K. Ganesan, R. Sarvendranath, and E. G. Larsson, "BeamSync: Over-the-air Synchronization for distributed massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 6824–6837, Jul. 2024.

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

[15] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility Maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[16] K. B. Petersen and M. S. Pedersen. "The matrix cookbook." 2012. [Online]. Available: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

[17] *Further Advancements for E-UTRA Physical Layer Aspects; (Release 9)*, 3GPP Standard TS 36.814, Mar. 2017.