

Adaptive Association for Cell-Free Massive MIMO Under Fronthaul Capacity Constraints

Ahmed AlAlwani^{*,||}, Karim Banawan^{*,†}, Mohammed Karmoose[†], Karim G. Seddik^{*}

^{*} Electronics and Communications Engineering Department, American University in Cairo, Egypt

^{||} Network Strategy & Engineering, Vodafone Intelligent Solutions (NSE VOIS), Smart Village, Giza, Egypt.

[†] Electrical Engineering Department, Faculty of Engineering, Alexandria University, Egypt.

E-mail: {aalalwani, karim.banawan, kseddik}@aucegypt.edu, mkarmoose@alexu.edu.eg

Abstract—Cell-free massive multiple-input multiple-output (CF-mMIMO) networks achieve high spectral efficiency through distributed, coherent transmission. However, strict fronthaul capacity constraints severely bottleneck practical deployments. Standard association algorithms often ignore these limitations, leading to overloaded access points (APs). This paper addresses fronthaul capacity constraints through three adaptive schemes. First, selective connection dropping identifies and eliminates redundant AP-user equipment (UE) links using a tunable priority score that allows operators to emphasize user redundancy, link quality, and/or load balancing based on their priorities. Second, hybrid decoding allows each user to decode signals from some APs jointly while treating others as separate streams, reducing the required fronthaul load. Third, SIC-enhanced hybrid decoding applies successive interference cancellation to further boost performance. We derive closed-form SINR expressions for each scheme and develop efficient iterative algorithms that share a common framework for identifying which AP-UE connections to adapt. Simulation results show that all three schemes achieve over 97% of the unconstrained baseline under relaxed fronthaul constraints, while preserving 60–84% under extremely tight constraints. Our results reveal that algorithm selection should be guided by the fronthaul capacity constraint severity and acceptable performance tradeoff.

Index Terms—Cell-free massive MIMO, user association, pilot assignment, user-centric CF-mMIMO, limited fronthaul.

I. INTRODUCTION

The escalating demand for higher data throughput and consistent quality of service (QoS) has exposed key limitations of traditional cell-centric architectures, such as inter-cell interference and degraded edge-user performance [1], [2]. Cell-free massive multiple-input multiple-output (CF-mMIMO) addresses these issues by eliminating cell boundaries and deploying numerous distributed access points (APs) connected to a central processing unit (CPU) to jointly serve user equipments (UEs) over shared time-frequency resources [3]. Early works considered fully cooperative architectures where every AP serves every UE [2]. However, this approach faces scalability issues due to excessive fronthaul signaling, global channel state information (CSI) requirements, and high computational complexity [4], [5]. Thus, recent works focus on user-centric (UC) architectures, where each UE is served by a subset of APs [6], [7]. This strategy reduces fronthaul traffic and enhances scalability. This motivates pursuing efficient AP-UE association and pilot assignment strategies.

Effective AP-UE association has therefore attracted significant attention, with numerous optimization and heuristic

approaches surveyed in [6]. A widely adopted baseline is the dynamic cooperation clustering (DCC) framework [5], where each UE is assigned a master AP and additional APs are assigned based on large-scale fading and pilot availability. However, while DCC enables scalable and stable clustering, it does not account for fronthaul limitations, often leading to excessive fronthaul load on heavily utilized APs. Building on this baseline, the coverage-aware preference-based resource association (CAPA) algorithm proposed in [8] improves sum spectral efficiency through iterative load balancing. In CAPA, UEs select preferred APs until a coverage threshold is met, while overloaded APs discard weak links to enforce capacity limits, followed by least-conflict pilot assignment (LCPA) to reduce pilot reuse among co-serving APs.

Beyond basic association, practical deployments face strict fronthaul capacity constraints. To address this, Femenias et al. [9] investigated max-min fairness resource allocation under strict fronthaul limits. The authors in [10] formulated UC clustering as a Mixed-Integer Linear Programming problem that minimizes the total number of AP-UE connections as a proxy for reducing fronthaul requirements without imposing explicit per-AP fronthaul-capacity constraints. Complementary research has focused on fronthaul management from the infrastructure side. E.g., dynamic AP-to-CPU clustering using the Hungarian algorithm has been proposed in [11]. In [12], an information-theoretic framework modeled physical fronthaul payloads using evolved Common Public Radio Interface (eCPRI) parameters to optimize joint user association and power allocation. Association complexity was reduced in [13] via hierarchical user grouping based on channel covariance similarity, while [14] addressed ultra-reliable low-latency communications (URLLC) applications by proposing rate-distortion-based fronthaul allocation to provide closed-form statistical QoS guarantees. While these works have advanced fronthaul-aware association and resource allocation, they assume coherent transmission where each UE's entire data stream must be forwarded by all its serving APs, consuming fronthaul capacity at each AP fronthaul link. Additionally, no prior work has studied the trade-off between performance, computational speed, and receiver complexity under varying fronthaul constraints. Our main contributions are:

- **Adaptive connection dropping with intelligent prioritization:** We propose a selective dropping algorithm to

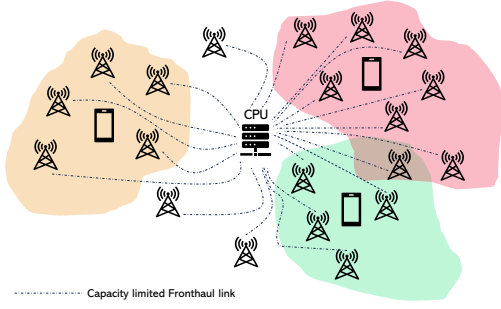


Fig. 1: User-centric cell-free massive MIMO network.

eliminate redundant AP-UE links using a novel tunable criterion that integrates user redundancy, AP importance, and multi-AP overload. Operators can balance spectral efficiency and convergence speed by adjusting the weight parameters based on network objectives.

- **Hybrid joint-separate decoding:** We introduce a hybrid scheme that partitions each UE's serving APs into two groups: those participating in joint transmission and those transmitting independently. The UE decodes these as a single joint stream and multiple separate streams, respectively, reducing the per-AP fronthaul load. We derive closed-form SINR expressions for each scheme.
- **SIC-enhanced hybrid decoding:** We extend the hybrid architecture with successive interference cancellation (SIC), where users first decode the joint stream, cancel it, and then sequentially decode and cancel the separate streams. This maintains unconstrained fronthaul performance even under severe fronthaul constraints.
- **Comprehensive performance analysis:** We evaluate the proposed schemes through complexity analysis and extensive simulations. Notably, our results demonstrate that the proposed algorithms consistently outperform all evaluated benchmarks. Furthermore, we quantify the trade-offs between performance and convergence speed, providing practical guidance for algorithm selection based on network constraints and receiver capabilities.

Notation: Bold lowercase $\mathbf{x}$ and uppercase $\mathbf{X}$ denote vectors and matrices, respectively. $\mathbf{X}^H$, $\mathbf{X}^{-1}$, and $\text{tr}(\mathbf{X})$ denote the Hermitian, inverse, and trace of $\mathbf{X}$, respectively. Calligraphic letters represent sets, with $|\mathcal{X}|$ as their cardinality. $\mathbf{I}_N$ and $\mathbf{0}_{N \times N}$ denote the $N \times N$ identity and zero matrices. $\mathbb{E}\{\mathbf{x}\}$ is the expectation of $\mathbf{x}$, and $\mathcal{CN}(\mu, \mathbf{R})$ denotes a complex Gaussian distribution with mean μ and covariance $\mathbf{R}$.

II. SYSTEM MODEL

We analyze a CF-mMIMO network comprising L geographically distributed APs each equipped with N antennas. The network operates in a coordinated manner to simultaneously serve K single-antenna UEs using shared time-frequency resources. The system operates under the massive MIMO regime where $L \times N \gg K$, ensuring that the aggregate number of antennas significantly exceeds the user population.

The channel between the k -th UE and ℓ -th AP is characterized by the channel vector $\mathbf{h}_{k\ell}$, which exhibits cor-

related Rayleigh fading statistics: $\mathbf{h}_{k\ell} \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{R}_{k\ell})$. Here, $\mathbf{R}_{k\ell} \in \mathbb{C}^{N \times N}$ denotes the spatial correlation matrix. We define the corresponding large-scale fading parameter as $\beta_{k\ell} = \text{tr}(\mathbf{R}_{k\ell})/N$. Due to sufficient spatial separation between APs (on the order of multiple wavelengths), we adopt the standard assumption of inter-AP channel independence: $\mathbb{E}\{\mathbf{h}_{k\ell}\mathbf{h}_{kj}^H\} = \mathbf{0}$ when $\ell \neq j$. The spatial correlation matrices $\mathbf{R}_{k\ell}$ are assumed known at the APs. Channel state information is acquired through time-division duplex (TDD) operation, exploiting UL/DL channel reciprocity.

We introduce $\mathcal{M}_k$ as the set of APs serving UE k , and $\mathcal{L}_\ell$ as the group of UEs served by AP ℓ . The association relationship is encoded via binary matrices $\mathbf{D}_{k\ell} \in \mathbb{C}^{N \times N}$:

$$\mathbf{D}_{k\ell} = \begin{cases} \mathbf{I}_N, & \text{if } \ell \in \mathcal{M}_k \\ \mathbf{0}_{N \times N}, & \text{otherwise} \end{cases} \quad (1)$$

Orthogonal pilot assignment across all UEs would eliminate pilot contamination. However, practical systems face constraints on the available pilot dimension τ_p . When $K > \tau_p$, full orthogonality is impossible, requiring pilot reuse among UEs. In the uplink (UL) training phase, UEs transmit their designated pilot sequences. The signal received at AP ℓ from UE k is denoted $\mathbf{y}_{k\ell}^p \in \mathbb{C}^{N \times 1}$. Applying minimum mean square error (MMSE) estimation [15], the channel estimate is given by $\hat{\mathbf{h}}_{k\ell} = \sqrt{\rho_p \tau_p} \mathbf{R}_{k\ell} \mathbf{\Psi}_{k\ell}^{-1} \mathbf{y}_{k\ell}^p$, where ρ_p represents the pilot transmission power, and the received signal covariance matrix is $\mathbf{\Psi}_{k\ell} = \sum_{i \in \mathcal{N}_k} \rho_p \tau_p \mathbf{R}_{i\ell} + \sigma^2 \mathbf{I}_N$ with σ^2 denoting noise variance and $\mathcal{N}_k$ denoting the subset of UEs sharing the same pilot. The estimation error is characterized by covariance matrix $\mathbf{C}_{k\ell} = \mathbf{R}_{k\ell} - \rho_p \tau_p \mathbf{R}_{k\ell} \mathbf{\Psi}_{k\ell}^{-1} \mathbf{R}_{k\ell}$. For downlink (DL) transmission, AP ℓ uses local partial-minimum mean square error (LP-MMSE) precoding to transmit to UE k . The precoding vector is computed as [15] $\bar{\mathbf{w}}_{k\ell} = \omega_{k\ell} \left(\sum_{i \in \mathcal{L}_\ell} \omega_{i\ell} (\hat{\mathbf{h}}_{i\ell} \hat{\mathbf{h}}_{i\ell}^H + \mathbf{C}_{i\ell}) + \sigma^2 \mathbf{I}_N \right)^{-1} \mathbf{D}_{k\ell} \hat{\mathbf{h}}_{k\ell}$ and $\mathbf{w}_{k\ell} = \bar{\mathbf{w}}_{k\ell} \cdot \frac{\sqrt{\omega_{k\ell}}}{\sqrt{\mathbb{E}[\|\bar{\mathbf{w}}_{k\ell}\|^2]}}$, where $\omega_{k\ell}$ is the power allocated by AP ℓ for transmitting to UE k given by $\omega_{k\ell} = \omega_{\max} \cdot \frac{(\beta_{k\ell})^v}{\sum_{i \in \mathcal{L}_\ell} (\beta_{i\ell})^v}$, where v controls the power allocation [15].

In the DL data transmission phase, all APs in $\mathcal{M}_k$ jointly transmit the same data symbol s_k intended for UE k , where $\mathbb{E}\{|s_k|^2\} = 1$ and $\mathbb{E}\{s_k s_i^*\} = 0$ for $k \neq i$. This enables joint coherent transmission, allowing each UE to combine signals from multiple APs and achieve array gain. While this establishes our baseline model, we will introduce modified transmission and decoding strategies later in the paper. For this initial setup, the received signal at UE k is given by $r_k = \sum_{\ell=1}^L \mathbf{h}_{k\ell}^H \mathbf{D}_{k\ell} \mathbf{w}_{k\ell} s_k + \sum_{i=1, i \neq k}^K \sum_{\ell=1}^L \mathbf{h}_{k\ell}^H \mathbf{D}_{i\ell} \mathbf{w}_{i\ell} s_i + n_k$. For compactness, we define the per-AP effective channel coefficient as $\Gamma_{ki}^{(\ell)} = \mathbf{h}_{k\ell}^H \mathbf{D}_{i\ell} \mathbf{w}_{i\ell}$, and the corresponding aggregate effective channel as $\Gamma_{ki} = \sum_{\ell=1}^L \Gamma_{ki}^{(\ell)}$. Here, Γ_{kk} represents the desired signal, while Γ_{ki} for $i \neq k$ is inter-user interference. Accordingly, the received signal can be written as $r_k = \Gamma_{kk} s_k + \sum_{i=1, i \neq k}^K \Gamma_{ki} s_i + n_k$. Since UE

k jointly decodes the signals received from all serving APs, the instantaneous DL SINR at UE k is then derived as:

$$\gamma_k = \frac{|\Gamma_{kk}|^2}{\sum_{i=1, i \neq k}^K |\Gamma_{ki}|^2 + \sigma_n^2} \quad (2)$$

The achievable DL spectral efficiency (SE) for UE k is then quantified as $\eta_k = \frac{\tau_d}{\tau_c} \log_2(1 + \gamma_k)$, where τ_c is the coherence block length and τ_d indicates the number of symbols allocated for DL data transmission within each coherence block.

The achievable SE expressions derived above assume ideal fronthaul links with unlimited capacity. In practical deployments, however, each AP is connected to the CPU through fronthaul links with limited capacity as shown in Fig. 1, which restricts the amount of downlink data that can be transported. The fronthaul load imposed on AP ℓ is determined by the aggregate SE of all UEs it serves, $C_\ell = \sum_{k \in \mathcal{L}_\ell} \eta_k$.

III. PROBLEM FORMULATION AND PROPOSED SOLUTION

We pose the problem of AP-UE association under fronthaul constraints as an aggregate DL SE maximization problem:

$$\begin{aligned} \max_{\{\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_L\}} \quad & \sum_{k=1}^K \eta_k \\ \text{s.t.} \quad & |M_k| \geq 1, \quad \forall k \end{aligned} \quad (3a)$$

$$|\mathcal{L}_\ell| \leq \tau_p, \quad \forall \ell \quad (3b)$$

$$\mathbf{D}_{k\ell} \in \{\mathbf{0}_{N \times N}, \mathbf{I}_N\}, \quad \forall k, \ell \quad (3c)$$

$$C_\ell \leq C_{\max}, \quad \forall \ell \quad (3d)$$

Constraint (3a) guarantees that each UE is assigned to at least one AP, while (3b) restricts the number of UEs served by any AP to be no greater than τ_p . Constraint (3c) represents the binary nature of the association decisions, and (3d) enforces the fronthaul capacity limitation. Due to its discrete and non-convex nature, finding the global optimum for this problem becomes computationally prohibitive. To address the fronthaul capacity constraints, we propose several heuristic approaches: (i) selectively dropping certain AP-UE connections while preserving others for each UE, (ii) transitioning from fully joint decoding to a hybrid scheme that combines joint decoding across a subset of APs with separate decoding from others, and (iii) employing SIC techniques to enhance SE within the fronthaul limitations. The scheme selection is adaptively determined based on the severity of the fronthaul constraint.

A. Selective Connection Dropping

When fronthaul capacity is limited, DCC often results in overloaded APs that violate constraint (3d). We address this through selective connection dropping that intelligently eliminates redundant links while preserving critical connections.

1) *Initialization and Adaptive Refinement*: We initialize the AP-UE association using the DCC algorithm [5]. To exploit short-term channel variations, we propose an adaptive refinement strategy that modifies the association pattern within each coherence block based on instantaneous channel realizations and fronthaul availability. Following the initial association, we identify the set of overloaded APs as those violating the fronthaul constraint $\mathcal{O} = \{\ell : C_\ell > C_{\max}\}$

2) *Connection Dropping Criterion*: The core challenge is determining which AP-UE connections to drop when an AP is overloaded. We only consider dropping connections from UEs with redundant links. Specifically, let $r_k = |\mathcal{M}_k|$ denote the number of APs serving UE k . The set of droppable candidates at overloaded AP ℓ is $\mathcal{U}_\ell = \{k \in \mathcal{L}_\ell : r_k > 1\}$, ensuring constraint (3a) remains satisfied after dropping.

For each UE $k \in \mathcal{U}_\ell$, we compute a dropping priority score $\alpha_{k\ell}$ that balances three objectives through tunable weights:

$$\alpha_{k\ell} = \tilde{r}_k^a \cdot (1 - I_{k\ell})^b \cdot \xi_k^c \quad (4)$$

where $a, b, c \in \{0, 1\}$ control emphasis on each factor. Higher $\alpha_{k\ell}$ indicates greater suitability for dropping. The factors are:

- **Normalized Redundancy**: $\tilde{r}_k = \frac{r_k - 1}{\max\{1, \max_i r_i - 1\}}$ measures excess connectivity beyond the minimum requirement of being connected to at least one AP. Higher values indicate the UE can better tolerate connection drops without service degradation.
- **AP Importance**: $I_{k\ell} = \frac{\beta_{k\ell}}{\sum_{m \in \mathcal{M}_k} \beta_{km}}$ quantifies the fraction of UE k 's total large-scale fading power from AP ℓ . Connections with lower importance contribute less to the UE's overall signal strength, making the term $(1 - I_{k\ell})$ high and the link more expendable.
- **Multi-Overload Penalty** (ξ_k): Let $\mathcal{O}_k = \mathcal{M}_k \cap \mathcal{O}$ denote the set of overloaded APs serving UE k . We define $\xi_k = \frac{|\mathcal{O}_k|}{\max\{1, |\mathcal{O}|\}}$. This factor is higher for UEs that contribute to the overload of multiple APs, making their connections attractive candidates for dropping.

The multiplicative structure ensures a connection is prioritized only when all three factors are favorable.

3) Iterative Overload Resolution and Residual Handling:

The connection dropping procedure, summarized in Algorithm 1, operates iteratively to alleviate fronthaul congestion. In each iteration, the algorithm identifies the most severely overloaded AP, ℓ^* , that serves at least one redundant UE ($r_k > 1$). From this AP, we break the connection to the UE k^* exhibiting the highest dropping priority score $\alpha_{k\ell^*}$. Following each drop, the LP-MMSE precoders, power allocations, and downlink SE are recomputed to reflect the updated network topology. This iterative reduction continues until either all APs satisfy the fronthaul limit, or dropping further connections would leave a UE completely unserved.

If the algorithm terminates but some APs remain overloaded (i.e., exclusively serving UEs with $r_k = 1$), connection dropping can no longer resolve the congestion. To enforce the fronthaul constraint for these cases without completely disconnecting users, we apply proportional SE scaling, which ensures that the total load on AP ℓ is capped at $C_{\max}$, while preserving proportional fairness among its served UEs:

$$\eta_k = (C_{\max}/C_\ell) \cdot \eta_k, \quad \forall k \in \mathcal{L}_\ell, \ell \in \mathcal{O} \quad (5)$$

B. Hybrid Joint-Separate Decoding

Instead of dropping connections to alleviate fronthaul congestion, we propose a hybrid strategy that retains all AP-UE

links but alters the transmission and decoding processes. We partition the serving APs for each UE k into two mutually exclusive sets: $\mathcal{B}_k$ for *joint decoding* and $\mathcal{J}_k$ for *separate decoding*. In conventional joint decoding, a UE with spectral efficiency η_k consumes the same rate η_k from the fronthaul links of all its serving APs. In contrast, when an AP $j \in \mathcal{J}_k$ transmits a separate symbol, it only consumes fronthaul proportional to the individual spectral efficiency contribution of that stream. Thus, moving selected APs from $\mathcal{B}_k$ to $\mathcal{J}_k$ reduces duplicated fronthaul consumption while preserving useful transmissions from multiple APs. Because APs in $\mathcal{J}_k$ transmit independent data streams, the received signal model is modified accordingly. Let s_k denote the common symbol transmitted by APs in $\mathcal{B}_k$, and let s_{kj} denote the independent symbol transmitted by AP $j \in \mathcal{J}_k$. Using the effective channel notation, the received signal at UE k can be expressed as:

$$r_k = \Gamma_{kk}^{(\mathcal{B}_k)} s_k + \sum_{j \in \mathcal{J}_k} \Gamma_{kk}^{(j)} s_{kj} + \sum_{i=1, i \neq k}^K \Gamma_{ki}^{(\mathcal{B}_i)} s_i + \sum_{i=1, i \neq k}^K \sum_{j \in \mathcal{J}_i} \Gamma_{ki}^{(j)} s_{ij} + n_k \quad (6)$$

where $\Gamma_{ki}^{(\mathcal{B}_i)} = \sum_{\ell \in \mathcal{B}_i} \Gamma_{ki}^{(\ell)}$. For the joint stream, UE k jointly decodes the signals received from $\mathcal{B}_k$. The resulting SINR is

$$\gamma_{\mathcal{B}_k} = \frac{|\Gamma_{kk}^{(\mathcal{B}_k)}|^2}{\sum_{i=1, i \neq k}^K |\Gamma_{ki}^{(\mathcal{B}_i)}|^2 + \sum_{i=1}^K \sum_{j \in \mathcal{J}_i} |\Gamma_{ki}^{(j)}|^2 + \sigma_n^2} \quad (7)$$

For each AP $\ell \in \mathcal{J}_k$, the receiver treats the signal as an independent stream with SINR

$$\gamma_k^\ell = \frac{|\Gamma_{kk}^{(\ell)}|^2}{\sum_{i=1}^K |\Gamma_{ki}^{(\mathcal{B}_i)}|^2 + \sum_{i=1}^K \sum_{j \in \mathcal{J}_i} |\Gamma_{ki}^{(j)}|^2 - |\Gamma_{kk}^{(\ell)}|^2 + \sigma_n^2} \quad (8)$$

The total achievable SE for UE k is then expressed as:

$$\eta_k = \frac{\tau_d}{\tau_c} \left[\log_2(1 + \gamma_{\mathcal{B}_k}) + \sum_{\ell \in \mathcal{J}_k} \log_2(1 + \gamma_k^\ell) \right] \quad (9)$$

To determine the optimal partition $\{\mathcal{B}_k, \mathcal{J}_k\}$, we initialize all connections as joint ($\mathcal{B}_k = \mathcal{M}_k, \mathcal{J}_k = \emptyset$). If initial load computations reveal overloaded APs, we iteratively migrate connections from joint to separate decoding following the same procedure as Algorithm 1. However, instead of dropping the connection (as in line 13 of Algorithm 1), the AP-UE link belonging to the UE k^* is shifted to $\mathcal{J}_{k^*}$. We then recompute the SE using (9) and update the AP loads, repeating this migration until all fronthaul constraints $C_\ell \leq C_{\max}$ are satisfied. If an AP remains overloaded even after all links in the system have been shifted to separate decoding, we apply the same proportional SE scaling used in the selective dropping approach to strictly enforce the fronthaul limit.

C. SIC-Enhanced Hybrid Decoding

The hybrid decoding scheme can be further improved by applying SIC at the receiver. The core principle is that interference from successfully decoded streams can be subtracted from the received signal before decoding subsequent ones. In this work, we assume ideal SIC, where previously decoded streams are perfectly recovered and cancelled without residual error. Specifically, once the joint stream is decoded, its interference is cancelled. Since it is decoded first, its SINR is unaffected $\gamma_{\mathcal{B}_k}^{\text{SIC}} = \gamma_{\mathcal{B}_k}$. Afterward, the separate streams are decoded sequentially. SIC reduces the effective interference for these later streams, enhancing their SINR and the overall SE. This performance improvement relies on the decoding order, typically processing streams from highest to lowest received signal power. Let $\mathcal{P}_k^\ell \subset \mathcal{J}_k$ be the set of separate streams already decoded and cancelled before processing stream ℓ . The SINR for stream $\ell \in \mathcal{J}_k$ then becomes:

$$\gamma_k^{\ell, \text{SIC}} = |\Gamma_{kk}^{(\ell)}|^2 / (\Lambda_k^{\ell, \text{SIC}} + \sigma_n^2) \quad (10)$$

where $\Lambda_k^{\ell, \text{SIC}}$ is the interference experienced by stream $\ell \in \mathcal{J}_k$ defined as:

$$\Lambda_k^{\ell, \text{SIC}} = \sum_{i=1, i \neq k}^K |\Gamma_{ki}^{(\mathcal{B}_i)}|^2 + \sum_{i=1}^K \sum_{j \in \mathcal{J}_i} |\Gamma_{ki}^{(j)}|^2 - |\Gamma_{kk}^{(\ell)}|^2 - \sum_{j \in \mathcal{P}_k^\ell} |\Gamma_{kk}^{(j)}|^2 \quad (11)$$

The combined SE remains in the same form as (9), but with improved SINR values due to SIC:

$$\eta_k^{\text{SIC}} = \frac{\tau_d}{\tau_c} \left[\log_2(1 + \gamma_{\mathcal{B}_k}^{\text{SIC}}) + \sum_{\ell \in \mathcal{J}_k} \log_2(1 + \gamma_k^{\ell, \text{SIC}}) \right] \quad (12)$$

Algorithm 1 Selective Connection Dropping

```

1: Input:  $\{\mathbf{D}_{k\ell}\}_{k=1, \ell=1}^{K, L}, C_{\max}, a, b, c$ 
2: Output:  $\{\mathbf{D}_{k\ell}\}_{k=1, \ell=1}^{K, L}, \{\eta_k\}_{k=1}^K, \{C_\ell\}_{\ell=1}^L$ 
3:  $\mathcal{O} \leftarrow \{\ell : C_\ell > C_{\max}\}$ 
4: while  $\mathcal{O} \neq \emptyset$  do
5:    $r_k \leftarrow |\mathcal{M}_k|, \forall k$ 
6:    $\mathcal{O}_{\text{red}} \leftarrow \{\ell \in \mathcal{O} : \exists k \in \mathcal{L}_\ell, r_k > 1\}$ 
7:   if  $\mathcal{O}_{\text{red}} = \emptyset$  then break
8:   end if
9:    $\ell^* \leftarrow \arg \max_{\ell \in \mathcal{O}_{\text{red}}} C_\ell$   $\triangleright$  Most overloaded AP
10:   $\mathcal{U}_{\ell^*} \leftarrow \{k \in \mathcal{L}_{\ell^*} : r_k > 1\}$   $\triangleright$  Droppable users
11:  Compute  $\alpha_{k\ell^*} = \tilde{r}_k^a (1 - I_{k\ell^*})^b \xi_k^c, \forall k \in \mathcal{U}_{\ell^*}$ 
12:   $k^* \leftarrow \arg \max_{k \in \mathcal{U}_{\ell^*}} \alpha_{k\ell^*}$ 
13:   $\mathbf{D}_{k^*\ell^*} \leftarrow \mathbf{0}_{N \times N}$ 
14:  Recompute  $\{\omega_{k\ell}\}, \{\mathbf{w}_{k\ell}\}, \{\eta_k\}, \{C_\ell\}$ 
15:  Update  $\mathcal{O}$ 
16: end while
17: for  $\ell \in \mathcal{O}$  do
18:    $\eta_k \leftarrow (\frac{C_{\max}}{C_\ell}) \eta_k, \forall k \in \mathcal{L}_\ell$ 
19: end for

```

TABLE I: Computational Complexity per Channel Realization

Algorithm	Algorithmic Complexity
DCC + SE Computation	$O(KL + K^2L)$
Selective Dropping	$O(KL + I_{drop} \cdot K^2L)$
Hybrid Decoding	$O(KL + I_{hybrid} \cdot K^2L)$
SIC-Enhanced Hybrid Decoding	$O(KL + I_{SIC} \cdot K^2L)$

D. Computational Complexity

We evaluate the computational complexity of the proposed schemes in terms of the number of UEs K , APs L , and iterations I , abstracting physical layer precoding complexity. As summarized in Table I, the baseline DCC association requires $O(KL)$ operations, while downlink SE computation scales as $O(K^2L)$. For the proposed adaptive algorithms, the complexity is dominated by the iterative SE recomputations. With a maximum of $I \leq KL - K$ iterations to resolve fronthaul congestion, the total asymptotic complexity scales as $O(I \cdot K^2L)$. Although all proposed schemes have similar network-side algorithmic complexity, they differ in *receiver implementation* requirements. Selective dropping relies on standard receivers, whereas the hybrid scheme requires dual decoding paths for joint and separate processing. The SIC-enhanced scheme incurs the highest hardware cost, requiring iterative cancellation, additional buffering, and increased latency. Thus, the performance gains of the hybrid and SIC approaches come at the expense of requiring more processing at the UE side rather than increased network-side complexity.

TABLE II: Convergence Comparison

Configuration	Iterations
Combined	17.6196
Importance	19.1073
Redundancy	15.9732
Overlap	12.4850

IV. PERFORMANCE EVALUATION

We evaluate four configurations of the tunable priority scoring function $\alpha_{k\ell} = \tilde{r}_k^a \cdot (1 - I_{k\ell})^b \cdot \xi_k^c$:

- **Combined Scoring** ($a = 1, b = 1, c = 1$): Balanced weighting of all three factors.
- **Importance-based** ($a = 0, b = 1, c = 0$): Prioritizes dropping weak connections based on AP importance.
- **Redundancy-based** ($a = 1, b = 0, c = 0$): Prefers dropping from well-connected users.
- **Overlap-based** ($a = 0, b = 0, c = 1$): Targets users contributing to multiple overloaded APs.

These configurations demonstrate the adaptability of the unified framework, with the last three representing special cases that isolate individual priority factors. For comparison, we include: **Random Dropping**, which randomly selects connections to be dropped from overloaded APs, and **Proportional Scaling**, which applies the scaling in (5) as the primary strategy instead of selective dropping.

We assess the proposed fronthaul management strategies through Monte Carlo simulations using a realistic CF-mMIMO network model based on the MATLAB implementation from [15] with the same channel parameters

specified in [Ch. 2, Sec. 5]. The network comprises $L = 30$ APs with $N = 4$ antennas, serving $K = 15$ single-antenna UEs. UEs are randomly distributed in a $1000 \text{ m} \times 1000 \text{ m}$ area with a 10 m vertical separation from APs. A wrap-around topology is employed to model an unbounded network¹. The system operates at 2 GHz with 20 MHz bandwidth. Each coherence interval has $\tau_c = 200$ symbols, with $\tau_p = 5$ for pilots. The maximum transmit power is 200 mW for APs in the DL and 100 mW for UEs in the UL.

Fig. 2 verifies that the proposed selective dropping scheme strictly enforces a fronthaul constraint of $C_{\max} = 25 \text{ bit/s/Hz}$. While the DCC approach which is oblivious to fronthaul limits allows AP loads to exceed 50 bit/s/Hz. The concentration of APs near the limit demonstrates efficient capacity utilization without violating infrastructure constraints. Next, we evaluate the effectiveness of the tunable priority score.

Fig. 3 plots the sum SE against the fronthaul capacity limit for the four weight configurations. Under relaxed constraints, all configurations perform near the unconstrained DCC baseline, confirming that redundant connections can be dropped without performance penalties. Under tight constraints, the Importance configuration excels, preserving 64% of the baseline SE, compared to 30.98% and 51.19% for the Redundancy/Overlap and Combined scoring approaches, respectively. This stems from its focus on preserving strong connections while eliminating weak ones, maintaining better overall channel quality even when forced to drop many connections. The Combined scoring scheme achieves moderate performance, whereas random dropping performs the worst, highlighting the necessity of intelligent prioritization.

To evaluate the convergence behavior, we compute the average number of iterations needed per-channel realization to solve the overload for each distinct weight configuration, as summarized in Table II. The results reveal a clear tradeoff: Importance achieves the highest SE but requires the most iterations, as focusing solely on link quality leads to conservative dropping. Conversely, Overlap converges fastest but yields the lowest SE by prioritizing load balancing without considering link quality. Combined Scoring strikes a favorable balance, requiring 7.78% fewer iterations than Importance while maintaining competitive performance. This makes it attractive for practical deployments where both performance and speed matter. For applications prioritizing convergence speed, Redundancy or Overlap offer 15-35% faster convergence at the cost of reduced SE. The framework's tunability allows operators to select weights based on their priorities.

Finally, Fig. 4 compares the proposed management strategies using the Importance configuration. The proportional scaling benchmark consistently yields the lowest performance with a substantial gap, highlighting the significant penalty of avoiding selective connection management entirely. At

¹In wrap-around topology, the northern boundary connects to the southern boundary, and the western boundary connects to the eastern boundary. This creates a toroidal geometry where all UEs experience interference from all directions, effectively simulating an infinite network [5].

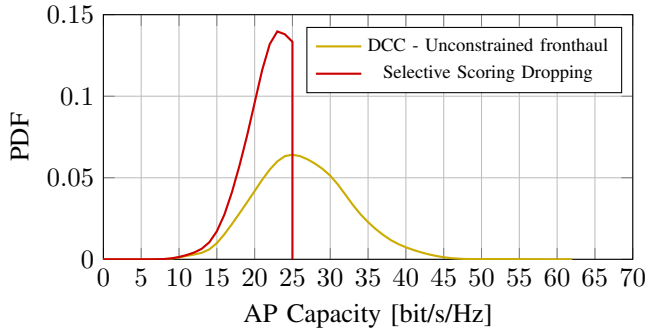


Fig. 2: AP load distribution with $C_{\max} = 25$ bit/s/Hz.

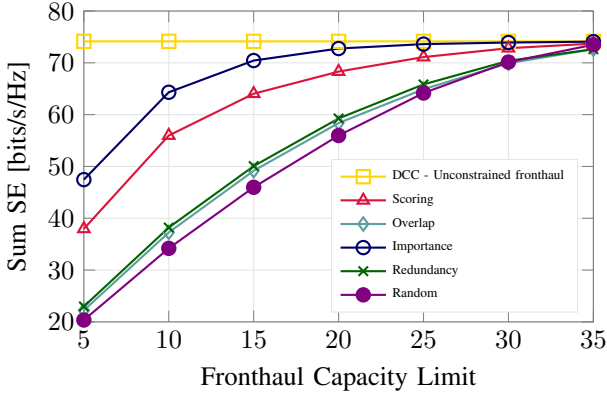


Fig. 3: Sum SE comparison of weight configurations.

relaxed constraints, all proposed schemes achieve over 97% of the DCC baseline, making the low-complexity selective dropping approach highly attractive. However, at extreme tight constraints, the SIC-enhanced hybrid scheme demonstrates its superiority reaching 83.89% of the baseline, outperforming pure selective dropping which achieves only 64% and standard hybrid decoding. Consequently, while simple dropping suffices for mild congestion, the SIC-enhanced architecture is essential for maintaining robust performance under tight fronthaul bottlenecks, justifying its increased receiver complexity.

V. CONCLUSION

This paper addressed fronthaul capacity constraints in CF-mMIMO systems through three adaptive strategies: se-

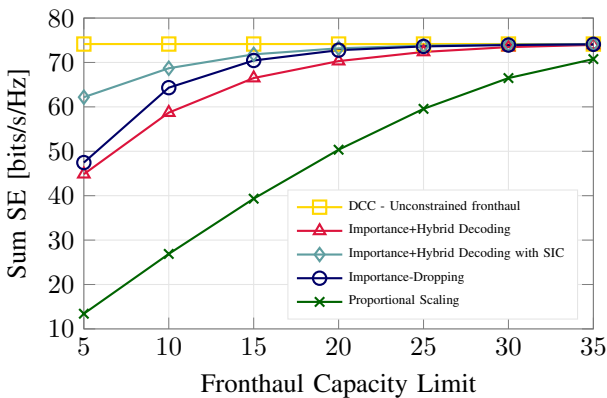


Fig. 4: Comparison of fronthaul management strategies.

lective connection dropping, hybrid joint-separate decoding, and SIC-enhanced hybrid decoding. By employing a tunable priority scoring function, our framework empowers operators to dynamically balance spectral efficiency and convergence speed. Evaluations revealed that while importance-based scoring maximizes overall throughput, a balanced weight configuration offers a practical compromise, converging 7.78% faster. Furthermore, under severe fronthaul bottlenecks, the SIC-enhanced hybrid scheme justifies its increased receiver complexity by effectively preserving near-baseline performance. Ultimately, the optimal management strategy is dictated by constraint severity, latency limits, and hardware capabilities. Future work might explore the dynamic optimization of the priority weighting exponents (a, b, c), potentially leveraging machine learning algorithms to adapt these values to time-varying network conditions and traffic demands.

REFERENCES

- [1] S. Vahid, R. Tafazolli, and M. Filo, "Small cells for 5G mobile networks," *Fundamentals of 5G Mobile Networks*, pp. 63–104, 2015.
- [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, Jan. 2017.
- [3] E. Nayeibi, A. Ashikhmin, T. L. Marzetta, and H. Yang, "Cell-free massive MIMO systems," in *IEEE Asilomar*, Nov. 2015, pp. 695–699.
- [4] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. IEEE ICC*, May 2019, pp. 1–6.
- [5] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. on Commun.*, vol. 68, no. 7, pp. 4247–4261, July 2020.
- [6] 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 Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 611–652, Dec. 2021.
- [7] S. Buzzi and C. D'Andrea, "Cell-Free Massive MIMO: User-Centric Approach," *IEEE Wireless Comm. Letters*, vol. 6, no. 6, pp. 706–709, Dec. 2017.
- [8] A. Alalwani, K. Banawan, K. G. Seddik, and M. Karmoose, "Scalable user association with least-conflict pilot assignment in cell-free massive MIMO," *IEEE Wireless Communications Letters*, vol. 15, pp. 1055–1059, Dec. 2025.
- [9] G. Femenias and F. Riera-Palou, "Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity," *IEEE Access*, vol. 7, pp. 44 596–44 612, April 2019.
- [10] C. F. Mendoza, S. Schwarz, and M. Rupp, "Cluster formation in scalable cell-free massive MIMO networks," in *2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, Oct. 2020, pp. 62–67.
- [11] M. Ajmal, A. Siddiqui, M. A. Tariq, M. M. Saad, and D. Kim, "Dynamic backhaul clustering for enhanced scalability in cell-free massive MIMO networks," in *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, April 2024, pp. 1735–1741.
- [12] Z. Mobini, A. H. Gokceoglu, L. Wang, G. Peters, and H. Q. Ngo, "Fronthaul-aware user-centric generalized cell-free massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 25, pp. 900–914, July 2025.
- [13] R. Wang, M. Shen, Y. He, and X. Liu, "Performance of cell-free massive MIMO with joint user clustering and access point selection," *IEEE Access*, vol. 9, pp. 40 860–40 870, Feb. 2021.
- [14] B. Chong, H. Lu, F. Guo, and Z. Xue, "Performance analysis for URLLC in cell-free massive MIMO systems with limited fronthaul capacity and hardware impairments," *IEEE Transactions on Wireless Communications*, vol. 25, pp. 163–176, June 2025.
- [15] Ö. T. Demir, E. Björnson, L. Sanguinetti *et al.*, "Foundations of user-centric cell-free massive MIMO," *Foundations and Trends® in Signal Processing*, vol. 14, no. 3-4, pp. 162–472, 2021.