# User-Centric Clustering Under Fairness Scheduling in Cell-Free Massive MIMO

Fabian Göttsch\*, Noboru Osawa<sup>†</sup>, Takeo Ohseki<sup>†</sup>, Yoshiaki Amano<sup>†</sup>,

Issei Kanno<sup>†</sup>, Kosuke Yamazaki<sup>†</sup>, Giuseppe Caire<sup>\*</sup>

\*Technical University of Berlin, Germany

<sup>†</sup>KDDI Research Inc., Japan

Emails: {fabian.goettsch, caire}@tu-berlin.de, {nb-oosawa, ohseki, yo-amano, is-kanno, ko-yamazaki}@kddi.jp

Abstract-We consider fairness scheduling in a user-centric cell-free massive MIMO network, where L remote radio units, each with M antennas, serve  $K \approx LM$  user equipments (UEs). Recent results show that the maximum network sum throughput is achieved where  $K_{\rm act} \approx \frac{LM}{2}$  UEs are simultaneously active in any given time-frequency slots. However, the number of users Kin the network is usually much larger. This requires that users are scheduled over the time-frequency resource and achieve a certain throughput rate as an average over the slots. We impose throughput fairness among UEs with a scheduling approach aiming to maximize a concave component-wise non-decreasing network utility function of the per-user throughput rates. In cell-free user-centric networks, the pilot and cluster assignment is usually done for a given set of active users. Combined with fairness scheduling, this requires pilot and cluster reassignment at each scheduling slot, involving an enormous overhead of control signaling exchange between network entities. We propose a fixed pilot and cluster assignment scheme (independent of the scheduling decisions), which outperforms the baseline method in terms of UE throughput, while requiring much less control information exchange between network entities.

*Index Terms*—User-Centric Clustering, Cell-Free Massive MIMO, Fairness Scheduling, Pilot Allocation.

### I. INTRODUCTION

Cell-free massive MIMO is a form of distributed massive MIMO that has attracted a great deal of interest in industry and research in recent years in order to serve a large number of user equipments (UEs) in dense beyond 5G networks. It is based on multiuser/massive MIMO [1]–[3], where the LM access point antennas are distributed across the network area on L remote radio units (RUs), each equipped with M antennas. A research direction towards a practical cell-free network considers a user-centric scalable system [4], [5], where user-centric clusters of RUs serve each UE  $k \in [K]$ .<sup>1</sup> Due to the distribution of RUs across the network area, cell-free massive MIMO is expected to serve all UEs with approximately the same quality of service. Unfortunately, this is not easy to achieve, even with efforts to make the network more fair [6]. While early works on cell-free massive MIMO assumed L > K [4, Ch. 2] leading to overall large UE data rates, recent works considered the more realistic UE density regime K > L, where K is comparable to ML (i.e., in the same order of magnitude) taking into account multi-antenna RUs. This regime yields a relatively unfair distribution of the UE data rates [6], [7].

For cell-free massive MIMO, very high-density scenarios are envisaged (e.g., see the real-world deployment in [8]). Since the total number of UEs K may be on the order of tens of thousands, it is clear that spatial multiplexing alone cannot support all UEs at the same time. Hence, for K significantly larger than ML, users must be scheduled in the time/frequency domain on different slots, such that on each "resource block" (RB), i.e., the slots defining the time-frequency granularity of the scheduler, only a number  $K_{\rm act}$  of "active" users is served using spatial multiplexing. In particular, recent results [9] have shown that for typical network layouts and operating conditions the network total spectral efficiency (SE) is maximum when  $K_{\rm act} \approx \frac{LM}{2}$ , and such maximum is quite "flat", i.e., quite insensitive with respect to the exact value of  $K_{\rm act}$ .

To give an idea, consider a network with 100 RBs per time slot serving K = 10000 users with L = 20 RUs and M = 16. Each user is allocated a block of F = 10 RBs in frequency to achieve a certain level of frequency diversity. Hence, a scheduler must choose on every RB a set of  $K_{\rm act} \approx 160$  users out of 1000 users per RB. Therefore, the relevant performance metric is the per-user throughput rate, i.e., the rate averaged over a long sequence of slots. Since the number of active users is much less than K, it is important to operate the network such that each user obtains a "fair share" of the total throughput rate. Hence, the scheduler must be designed to achieve some desired form of fairness of the per-user throughput distribution. Finally, also as a consequence of this setting, we notice that the familiar ergodic rate used as performance metric in most standard literature on cell-free networks (e.g., see [4]–[7]) is not relevant any longer. In fact, the scheduler must allocate an "instantaneous" rate on each RB (or block of F RBs) and decoding is performed block by block, such that averaging over a virtually infinite sequence of fading states is no longer possible. In this case, the instantaneous rate must be scheduled according to the notion of information outage rate (e.g., see [10]).

### A. Related Literature

Scheduling in cell-free massive MIMO has been considered in relatively few works [9], [11]–[13] in comparison with the very large number of papers considering the ergodic rate of a fixed set of "always active" users (see, e.g., [4] and references therein). We build on the system proposed in [9], which considers  $K > K_{act}$  UEs, and sets  $K_{act}$  as the number

<sup>&</sup>lt;sup>1</sup>The set of integers from 1 to N is denoted by [N].

of active users that (approximately) maximizes the network total SE. Using the dynamic scheduling framework of [14], hard and proportional fairness scheduling (resp., HFS and PFS) are addressed. A system design challenge with a high UE density consists of the assignment of uplink (UL) pilots for channel estimation and the user-centric cluster formation [4], [7]. This problem is exacerbated in the presence of dynamic scheduling, since the set of active user changes at every scheduling decision (time slot). In [9], this challenge is addressed by performing the pilot and cluster assignment at each scheduling decision, i.e., after selecting the set of  $K_{\rm act} < K$  active UEs. In fact, users sharing the same pilot and located in proximity of each other may suffer from severe pilot contamination. Nevertheless, pilot contamination comes only from active users, the identity of which is not known prior the scheduling decision on the current slot. In practice, performing these operations at each slot comes at the cost of a large communication overhead between UEs and RUs.

#### B. Contributions

We propose a less communication-intensive pilot and cluster assignment scheme with pilots and clusters permanently assigned to the K UEs. For those co-pilot users that may cause severe mutual pilot contamination, we construct a conflict graph that prohibits these UEs from being scheduled on the same RB. The proposed method requires far less communication and it is much better suited for a practical implementation, compared to the per-slot reassignment scheme in [9]. Interestingly, our simulations show that it can also achieve a (slightly) larger throughput rate, due to the avoidance of strong co-pilot interference expressed by the conflict graph.

Furthermore, as anticipated before, we use information outage rates for the instantaneous rate scheduling, reflecting the fact that channel coding is performed on a block by block basis on a finite number of RBs. In particular, our numerical results show the benefit of a moderate frequency diversity order of F > 1 RBs. As F increases, the instantaneous mutual information distribution "concentrates" and behaves in a more deterministic way, allowing a more aggressive instantaneous rate allocation on the active slots.<sup>2</sup>

#### **II. SYSTEM DESCRIPTION**

We consider a cell-free massive MIMO network as in [9] in TDD operation mode with L RUs, each equipped with Mantennas, and K single-antenna UEs. Both RUs and UEs are distributed on a squared region on the 2-dimensional plane. We let  $\mathbb{H}(t, f) \in \mathbb{C}^{LM \times K}$  denote the channel matrix between all the K UE antennas and all the LM RU antennas on a given RB f in time slot t, formed by  $M \times 1$  blocks  $\mathbf{h}_{\ell,k}(t, f)$  in correspondence of the M antennas of RU  $\ell$  and UE k. Letting  $\mathcal{A}(t) \subseteq [K]$  denote the set of active users scheduled in slot t, the columns  $h_k(t, f)$  of  $\mathbb{H}(t, f)$  corresponding to inactive UEs  $k \in [K] : k \notin \mathcal{A}(t)$  contain the identically zero vector **0**.

Let **F** denote the  $M \times M$  unitary DFT matrix with (m, n)elements  $[\mathbf{F}]_{m,n} = \frac{e^{-j\frac{2\pi}{M}mn}}{\sqrt{M}}$  for  $m, n = 0, 1, \dots, M-1$ , and consider the angular support set  $S_{\ell,k} \subseteq \{0, \dots, M-1\}$ obtained according to the single ring local scattering model [15], where  $S_{\ell,k}$  contains the DFT quantized angles (multiples of  $2\pi/M$ ) falling inside an interval of length  $\Delta$  placed symmetrically around the direction joining UE k and RU  $\ell$ . Then, the channel between RU  $\ell$  and the active UE k with large-scale fading coefficient (LSFC)  $\beta_{\ell,k}$  on RB f in slot t is  $\mathbf{h}_{\ell,k}(t,f) = \sqrt{\frac{\beta_{\ell,k}M}{|S_{\ell,k}|}} \mathbf{F}_{\ell,k} \boldsymbol{\nu}_{\ell,k}(t,f)$ , where, using a MATLAB-like notation,  $\mathbf{F}_{\ell,k} \stackrel{\Delta}{=} \mathbf{F}(:, \mathcal{S}_{\ell,k})$  denotes the tall unitary matrix obtained by selecting the columns of  $\mathbf{F}$ corresponding to the index set  $S_{\ell,k}^3$ , and  $\nu_{\ell,k}(t,f)$  is an  $|\mathcal{S}_{\ell,k}| \times 1$  i.i.d. Gaussian vector with components ~  $\mathcal{CN}(0,1)$ . Note that the LSFC, angular support and thus also  $\mathbf{F}_{\ell,k}$  are independent of the RB and time indices, while  $\nu_{\ell,k}(t, f)$  has different realizations on each RB f and in each slot t. As in most cell-free massive MIMO literature (see [4]), we assume that the LSFCs are known at the RUs.

In slot t and for all RBs  $f \in [F]$ , each UE k is connected to a cluster  $C_k(t) \subseteq [L]$  of RUs and each RU  $\ell$  has a set of associated UEs  $\mathcal{U}_{\ell}(t) \subseteq [K]$ . The UE-RU association is described by a bipartite graph  $\mathcal{G}(t)$  with two classes of nodes (UEs and RUs) such that the neighborhood of UE-node k is  $C_k(t)$  and the neighborhood of RU-node  $\ell$  is  $\mathcal{U}_{\ell}(t)$ . The set of edges of  $\mathcal{G}(t)$  is denoted by  $\mathcal{E}(t)$ , i.e.,  $\mathcal{G}(t) = \mathcal{G}([L], [K], \mathcal{E}(t))$ . We assume OFDM modulation and that the channel in the timefrequency domain follows the standard block-fading model [3]-[5]. The channel vectors from UEs to RUs are random but constant over coherence blocks of T signal dimensions in the time-frequency domain, of which  $\tau_p$  dimensions are used for the finite-dimensional UL pilot signal, such that a fraction  $1 - \frac{\tau_p}{T}$  of dimensions per RB is used for data transmission. We assume that one time-frequency slot, i.e., one realization of t and f, corresponds to one channel coherence block.

## A. Uplink Decoding with Partial Channel State Information

We consider *partial* channel state information obtained by subspace projection channel estimation. Each RU  $\ell$  computes locally the channel estimates  $\hat{\mathbf{h}}_{\ell,k}(t, f)$  for UEs  $k \in \mathcal{U}_{\ell}(t)$ from the received *orthogonal* UL pilot signal, where perfect subspace knowledge is assumed (see [7] for details).

Based on the channel estimates  $\{\mathbf{h}_{\ell,k}(t, f) : k \in \mathcal{U}_{\ell}(t)\}$ , RU  $\ell$  locally computes a unique receiver combining vector  $\mathbf{v}_{\ell,k}(t, f)$  for each associated UE  $k \in \mathcal{U}_{\ell}$ , where a linear MMSE principle is used. For  $k \notin \mathcal{U}_{\ell}(t)$ , we have  $\mathbf{v}_{\ell,k}(t, f) =$ **0**. The cluster  $\mathcal{C}_k(t)$  combines the vectors  $\{\mathbf{v}_{\ell,k}(t, f) : \ell \in \mathcal{C}_k(t)\}$  to form a receiver *unit norm* vector  $\mathbf{v}_k(t, f) \in \mathbb{C}^{LM \times 1}$ 

 $<sup>^{2}</sup>$ We consider scheduling over time and allocate all *F* RBs to active users. The allocation of RBs to different users in the same time slot to avoid conflicts is beyond the scope of this work. The proposed conflict graph-based scheme would be an approach to assign users to different RBs if it is done in a sequential manner, e.g., RB by RB. The problem would then be very similar to what is done in this paper.

<sup>&</sup>lt;sup>3</sup>Note that for uniform linear arrays (ULAs) and uniform planar arrays (UPAs), as widely used in today's massive MIMO implementations, the channel covariance matrix is Toeplitz (for ULA) or Block-Toeplitz (for UPA), and that large Toeplitz and block-Toeplitz matrices are approximately diagonalized by DFTs on the columns and on the rows (see [15] for a precise statement based on Szegö's theorem).

for UE k, aiming to maximize the UL signal to Interference plus noise ratio (SINR) (see [16] for details). Note that the cluster combining uses weights to fuse the signals from the RUs  $\ell \in C_k(t)$ , which replace power control in the UL [4, Sec. 2.6]. It is further shown in [4, Sec. 7.3] that uniform UL power allocation yields comparable results compared to common power control schemes in cell-free networks. We focus on UL results, since by duality, the UL and downlink data rates and system performance are almost identical [5], [17], [18].

## III. UPLINK DATA TRANSMISSION

Let all active UEs transmit with the same average energy per symbol  $P^{\text{ue}}$ , and we define the system parameter  $\text{SNR} \stackrel{\Delta}{=} P^{\text{ue}}/N_0$ , where  $N_0$  denotes the complex baseband noise power spectral density. The received  $LM \times 1$  symbol vector at the LM RU antennas for a single channel use on RB f in slot tof the UL is given by

$$\mathbf{y}(t,f) = \sqrt{\mathsf{SNR}} \,\mathbb{H}(t,f)\mathbf{s}(t,f) + \mathbf{z}(t,f),\tag{1}$$

where  $\mathfrak{s}(t, f) \in \mathbb{C}^{K \times 1}$  is the vector of information symbols transmitted by the UEs on RB f in slot t (zero-mean unit variance and mutually independent random variables) and  $\mathbb{Z}(t, f)$  is an i.i.d. noise vector with components  $\sim \mathcal{CN}(0, 1)$ . The goal of cluster  $\mathcal{C}_k(t)$  is to produce an effective channel observation for symbol  $s_k(t, f)$ , the k-th component of the vector  $\mathfrak{s}(t, f)$ , from the collectively received signal at the RUs  $\ell \in \mathcal{C}_k(t)$ . Using the receiver vector  $\mathbb{V}_k(t, f)$ , the corresponding scalar combined observation for symbol  $s_k(t, f)$  is given by  $\hat{s}_k(t, f) = \mathbb{V}_k(t, f)^{\mathsf{H}}\mathfrak{Y}(t, f)$ . We let  $\mathbb{H}(t) \triangleq \mathbb{H}(t, 1 : F)$ and  $\mathbb{V}_k(t) \triangleq \mathbb{V}_k(t, 1 : F)$  denote the realization of the channel matrix and of the receiver vector for UE k in time slot t for RBs  $f = \{1, \ldots, F\}$ , respectively. The instantaneous mutual information  $I(\{\hat{s}_k(t, f) : f \in [1 : F]\}; \{s_k(t, f) : f \in [1 : F]\})$  in slot t is a function of  $\{\mathbb{V}_k(t), \mathbb{H}(t)\}$  and given by<sup>4</sup>

$$\mathcal{I}_k(\mathbb{v}_k(t), \mathbb{H}(t)) \triangleq \frac{1}{F} \sum_{f=1}^F \log(1 + \mathsf{SINR}_k(t, f)), \quad (2)$$

where we define

$$\mathsf{SINR}_{k}(t,f) = \frac{|\mathbb{v}_{k}(t,f)^{\mathsf{H}}\mathbb{h}_{k}(t,f)|^{2}}{\mathsf{SNR}^{-1} + \sum_{j \neq k} |\mathbb{v}_{k}(t,f)^{\mathsf{H}}\mathbb{h}_{j}(t,f)|^{2}}.$$
 (3)

### A. Rate Allocation

Following [9], we consider outage rates as the effective data rates, such that the receiver can reliably decode an allocated rate under the condition that no information outage occurs [19]. This condition holds if the allocated rate  $r_k$  is smaller than the mutual information  $\mathcal{I}_k(\mathbb{v}_k(t), \mathbb{H}(t))$ . The effective instantaneous service rate of UE k in time slot t (expressed in bit per time-frequency channel use) is thus given by [10]

$$\mu_k(t) = \begin{cases} (1 - \frac{\tau_p}{T})R_k(t), & \text{if } k \in \mathcal{A}(t), \\ 0, & \text{if } k \notin \mathcal{A}(t), \end{cases}$$
(4)

where, letting  $\mathbb{1}{S}$  be the indicator function of an event S,

$$R_k(t) \triangleq r_k \times \mathbb{1}\{r_k < \mathcal{I}_k(\mathbb{v}_k(t), \mathbb{H}(t))\}.$$
 (5)

Notice that in the information outage regime, even if a user is active (i.e.,  $k \in \mathcal{A}(t)$ ), it may still have zero rate, depending on the condition  $\{r_k < \mathcal{I}_k(\mathbb{V}_k(t), \mathbb{H}(t))\}$ . In fact, while the channel state information may be assumed known at the receiver, it is definitely not known at the transmitter, such that instantaneous slot-by-slot rate allocation is not possible. Instead,  $r_k$  must be chosen on the basis of the random variable  $\mathcal{I}_k(\mathbb{V}_k(t),\mathbb{H}(t))$ . In stationary conditions, the instantaneous mutual information distribution is independent of the slot time t. In practice, with moderate user mobility, this changes slowly over time. In addition, it is very difficult to analytically characterize such distribution since in general  $SINR_k(t, f)$  in (3) depends not only on the channel state, but also on the set of active users. For the time being, we assume such distribution to be known for each user k. Later in this section we present an adaptive algorithmic solution for effective rate allocation.

The per-user throughput rate is defined as

$$\bar{\mu}_k = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mu_k(\tau) = \mathbb{E}[\mu_k(\mathbb{H})], \tag{6}$$

where, with some abuse of notation, we denote by  $\mu_k(\mathbb{H})$  the random variable induced by the scheduling policy (i.e., the selection of  $r_k$  and the active set  $\mathcal{A}(t)$  in (4) as a function of the channel state  $\mathbb{H}$ ), and where the convergence of the time average in (6) is guaranteed by the ergodicity of the channel process (in our case, i.i.d. over the RBs) and the stationarity of the scheduling policy in the class of dynamic policies considered here [14].

Letting the complementary cumulative distribution function (CDF) of the instantaneous mutual information of user k be  $P_k(z) \triangleq \mathbb{P}(\mathcal{I}_k(\mathbb{v}_k, \mathbb{H}) > z)$ , from (5) we have that  $\mathbb{E}[R_k(t)] = r_k P_k(r_k)$ . Hence, the optimization of  $r_k$  is immediate and yields [10]

$$r_k^* = \arg\max \ z \times P_k(z). \tag{7}$$

Since, as said before, the statistics of  $\mathcal{I}_k(\mathbb{v}_k(t), \mathbb{H}(t))$  for a UE k are generally extremely hard to obtain and depend on the scheduling policy itself, here we consider the localized adaptive scheme proposed in [9], where each user k collects a sliding window of N past samples of the instantaneous mutual information and optimizes its transmission rate  $r_k$  using the empirical complementary CDF based on these samples.<sup>5</sup> In the following, we let  $\bar{R}_k \triangleq r_k^* P_k(r_k^*)$ , i.e., the maximum of the objective function in the right-hand side of (7).

## IV. SYSTEM OPTIMIZATION

We consider a network in the UL with a total number of K UEs, which operates at its optimal load, when serving  $K_{\text{act}} < K$  UEs. Further, we assume an *infinite backlog* situation, where each UEs has an infinite buffer of data to transmit.

<sup>&</sup>lt;sup>4</sup>This is in the assumption that the channel state information is known at the receiver and "Gaussian" single-user codebooks are used.

<sup>&</sup>lt;sup>5</sup>As in [9], the allocated rates are initialized by a "start-up" phase consisting of  $N_{\text{init}}$  time slots. In each of the  $N_{\text{init}}$  time slots  $K_{\text{act}}$  out of the K UEs are, considering the conflict graph, randomly selected to be active. In practice, a user joining the system would start with a very conservative rate and progressively "ramp up" the value of  $r_k$  until the maximum of the product in (7) is achieved. Actual practical algorithms for rate scheduling work on averaged local statistics along these lines, such that non-stationary (slowly varying) statistics can be tracked.

By scheduling at most  $K_{\rm act}$  UEs per time slot, the scheduler wishes to maximize the network utility function, defined as a suitable concave component-wise non-decreasing function  $g(\cdot)$  of the user throughput rate vector  $\bar{\boldsymbol{\mu}} = (\bar{\mu}_1, \dots, \bar{\mu}_K)$ . The problem to be solved is

maximize 
$$g(\bar{\mu})$$
, subject to  $\bar{\mu} \in \mathcal{R}$ , (8)

where  $\mathcal{R}$  is the system achievable throughput rate region [14]. Since  $\mathcal{R}$  is not characterized easily, the solution  $\bar{\mu}^*$  of (8) is generally very hard to find analytically [10]. However, the framework of [14] can be used to find a scheduling scheme that approximates  $\bar{\mu}^*$  within any desired accuracy.

Specifically, the scheduler solves at each scheduling slot t the weighted sum rate maximization (WSRM) problem (with respect to the active set A(t))

$$\max \sum_{k \in \mathcal{A}(t)} Q_k(t) \bar{R}_k, \tag{9}$$

where  $\{Q_k(t)\}\$  are the backlogs of "virtual queues" used as weights in (9) with update rule

$$Q_k(t+1) = \max\{Q_k(t) - \mu_k(t), 0\} + A_k(t)$$
(10)

and  $\{A_k(t)\}\$  is a set of "virtual arrival processes". For each t, we have  $A_k(t) = a_k$ , where  $\mathbf{a} = (a_1, \dots, a_K)$  is the solution to the convex optimization problem

$$\begin{array}{ll} \underset{\mathbf{a}}{\operatorname{maximize}} & Vg(\mathbf{a}) - \sum_{k \in [K]} Q_k(t) a_k \\ \text{subject to} & 0 \leq a_k \leq A_{\max}, \quad \forall k \in [K]. \end{array}$$
(11)

Here, V and  $A_{\text{max}}$  are suitably chosen constant parameters that determine the behavior of the algorithm [14]. In particular, it is known that for  $A_{\text{max}}$  sufficiently large the time-averaged service rates generated by the algorithm (i.e.,  $\frac{1}{t} \sum_{\tau=1}^{t} \mu_k(\tau)$ for large t) approximate the optimal throughput rate point solution of (8) within a gap O(1/V), while the time-averaged sum queue backlog grows as O(V).<sup>6</sup>

# A. Proportional Fairness and Hard Fairness Scheduling

We consider PFS and HFS, leading to different solutions to the optimization problem (11). In case of PFS, we have  $g(\mathbf{a}) = \sum_{k \in [K]} \log a_k$  in (11), which yields the arrivals [10]

$$a_k = \min\left\{\frac{V}{Q_k(t)}, A_{\max}\right\}.$$
 (12)

For HFS,  $g(\mathbf{a}) = \min_{k \in [K]} a_k$  and the solution to (11) is [10]

$$a_k = \begin{cases} A_{\max}, & \text{if } V > \sum_{k \in [K]} Q_k(t), \\ 0, & \text{else.} \end{cases}$$
(13)

# V. ALGORITHMIC SOLUTIONS

We first describe an algorithmic solution proposed in [9] including a reassignment of pilots and clusters at each scheduling decision. Then we will describe our proposed scheme with fixed pilot and cluster assignments, reducing greatly the required communication between UEs and RUs. For both schemes,  $\mathbf{Q}(0) = \mathbf{0}$ , and UEs with empty queues are not scheduled, so in some slots the number of active UEs may be smaller than  $K_{\text{act}}$ , in particular when HFS is employed.

#### A. Pilot and Cluster Reallocation Scheme

This scheme proposed in [9] carries out the UL pilot allocation for channel estimation and cluster formation in each time slot after selecting the set of active UEs. Having defined  $K_{\text{act}}$  as the desired number of simultaneously active UEs, we solve the WSRM (9) as

$$\begin{array}{ll} \underset{\mathbf{x}}{\operatorname{maximize}} & \sum_{k \in [K]} Q_k(t) \bar{R}_k x_k \\ \text{subject to} & \sum_{k \in [K]} x_k \leqslant K_{\operatorname{act}}, \\ & x_k \in \{0, 1\}, \end{array}$$
(14)

where  $x_k = 1$  if UE  $k \in \mathcal{A}(t)$  and 0 otherwise. The solution is immediate and consists of sorting the users in decreasing order of the product  $Q_k(t)\overline{R}_k$  and letting  $\mathcal{A}(t)$  the set of the top  $K_{\text{act}}$  sorted users. Given the selected set  $\mathcal{A}(t)$ , UL pilots and user-centric clusters are assigned to the active UEs following the semi-overloaded pilot assignment method from [20], where an RU may assign the same pilot sequence to multiple UEs provided that the channel subspaces of the UEs are nearly mutually orthogonal, such that accurate channel estimation is possible (negligible mutual pilot contamination, using the decontamination method of [7]). An RU-UE association can only be established, when the SNR association threshold criterion  $\beta_{\ell,k} \geq \frac{\eta}{M \text{SNR}}$  is fulfilled, where  $\eta$  is an association threshold parameter.

# B. Fixed Pilots and Clusters

In this case, we first define a conflict graph  $\mathcal{C} = ([K], \mathcal{E}_{\mathcal{C}})$ with a vertex set corresponding to all K UEs in the network and an edge set  $\mathcal{E}_{\mathcal{C}}$  accounting for the conflicts. Letting  $p_k$ denote the UL pilot index of UE k, we define that a UE-pair (k, k') has a scheduling conflict if

- 1) the UEs are associated to at least one common RU, i.e.,  $C_{k,k'} \triangleq C_k \cap C_{k'} \neq \emptyset$ , and
- 2) the UEs are assigned the same UL pilot, i.e.,  $p_k = p_{k'}$ , and
- the subspaces of the UEs overlap with regard to at least one RU ℓ ∈ C<sub>k,k'</sub>, i.e., ∑<sub>ℓ∈C<sub>k,k'</sub> |S<sub>ℓ,k</sub> ∩ S<sub>ℓ,k'</sub> | > 0, where | · | denotes the cardinality of a set.
  </sub>

The graph  $\mathcal{C}$  has an edge between the vertex k and vertex k' for all UE-pairs (k, k') in conflict, with the meaning that any UE-pair  $(k, k') \in \mathcal{E}_{\mathcal{C}}$  is not allowed to be scheduled in the same time slot, since they would interfere in the channel estimation process. Based on this conflict definition, we propose the following pilot assignment and cluster formation scheme.

- 1) When a UE k joins the system, it connects to a maximum of Q RUs with the largest LSFCs, provided that  $\beta_{\ell,k} \ge \frac{\eta}{M \text{SNR}}$ , forming the set  $C_k$ .<sup>7</sup>
- 2) For each UL pilot index  $\tau^{(i)} = [\tau_p]$  the RUs  $\ell \in C_k$  count the number of associated UEs  $k' \neq k : k' \in U_\ell$  with a

 $^{7}$ In a practical system, UEs join and leave the network, such that each UE could be assigned a pilot and an RU cluster according to the proposed scheme. In our simulations, we carry out the proposed scheme for each UE in the order of their index.

<sup>&</sup>lt;sup>6</sup>The proof under the assumptions made in this paper differs from the performance guarantees given in [10], [14] and will be published in a journal paper on scheduling in cell-free massive MIMO. Further, in most literature, the timeaveraged queue backlog is referred to as "delay" but here since we are in the infinite buffer regime and the queues are virtual, this quantity is rather an indication of the time it takes for the algorithm to converge to a stationary state.



Fig. 1. User throughput for HFS (left) and PFS (right), where "Baseline" accounts for the reassignment scheme from [9].

non-orthogonal subspace. The set of potentially conflicting UEs with pilot  $\tau^{(i)}$  regarding UE k is given by  $\mathcal{C}_k(\tau^{(i)}) = \{\bigcup_{\ell \in \mathcal{C}_k} \mathcal{C}_{\ell,k}(\tau^{(i)})\}$ , where

$$\mathcal{C}_{\ell,k}(\tau^{(i)}) = \{k' \in \mathcal{U}_{\ell} : \mathbb{1}\{p_{k'} = \tau^{(i)}\} \cap \mathbb{1}\{|\mathcal{S}_{\ell,k} \cap \mathcal{S}_{\ell,k'}| > 0\}\},\$$

and where  $C_k(\tau^{(i)})$  contains each possible UE only once. Here, perfect subspace knowledge at RU  $\ell$  for associated UEs  $k \in U_\ell$  is assumed. Schemes for channel subspace and covariance matrix estimation, respectively, in cell-free massive MIMO are presented in [7], [21].

The pilot corresponding to the smallest number of conflicting UEs, i.e., τ<sup>(i\*)</sup> = arg min |C<sub>k</sub>(τ<sup>(i)</sup>)|, is assigned to UE k. If more than one pilot corresponds to the smallest number of conflicting UEs, an arbitrary choice of these pilots is made.

The fixed pilot and cluster assignment to all K UEs in the network is carried out independently of scheduling decisions. The resulting WSRM problem (9), subject to the conflict graph, is given by the linear integer program

$$\begin{array}{ll} \underset{\mathbf{x}}{\operatorname{maximize}} & \sum_{k \in [K]} Q_k(t) \bar{R}_k x_k \\ \text{subject to} & \sum_{k \in [K]} x_k \leqslant K_{\operatorname{act}}, \\ & x_k \in \{0, 1\}, \\ & x_k + x_{k'} \leqslant 1, \ \forall (k, k') \in \mathcal{E}_{\mathfrak{C}}, \end{array}$$
(15)

which can be efficiently solved with standard tools (e.g., Gurobi or MATLAB) even for fairly large systems.

## VI. NUMERICAL EVALUATIONS AND OUTLOOK

We consider a system like in [9], i.e., a squared network area of  $A = 50 \times 50 \text{m}^2$  with a torus topology to avoid boundary effects, containing L = 12 RUs, each with M = 8antennas, and K = 100 UEs. We assume a bandwidth of W = 10 MHz and noise with power spectral density of  $N_0 = -174$  dBm/Hz. We let the angular interval of length  $\Delta = \pi/8$ , the SNR threshold  $\eta = 1$  and the maximum cluster size Q = 10 (RUs serving one UE) in the simulations. We define  $P^{ue}$  such that  $\bar{\beta}MSNR = 1$  (i.e., 0 dB), when the expected pathloss  $\bar{\beta}$  with respect to LOS and NLOS is calculated for distance  $3d_L$ , where  $d_L = \sqrt{\frac{A}{\pi L}}$  is the radius of a disk of area equal to A/L. This leads to a certain level of overlap of the RUs' coverage areas considering the SNR association threshold. The UEs are randomly dropped in the network area, while the RUs are placed on  $3 \times 4$  rectangular grid. The online rate adaptation is carried out for all schemes with  $N_{\text{init}} = 500$  and N = 100, and we consider RBs of dimension T = 200 symbols. Since we consider a network like



Fig. 2. The empirical CDF of  $\mathcal{I}_k(\mathbb{V}_k, \mathbb{H})$  of an example UE k (left) and the user throughput (right) with  $F = \{1, 8\}$  in a network employing PFS.

in [9], we choose  $\tau_p = 20$  and  $K_{\text{act}} = 40$ , the approximate optimal parameters according to the results in [9]. We simulate 5 different setups (random placement of UEs) that are equal for all investigated approaches. The algorithm parameters are chosen as  $A_{\text{max}} = 100$  and V = 10000. We refer the reader to [9], [10] for an evaluation of different values of V.

## A. Fixed Pilots and Clusters vs. Reassignment Scheme

Considering a narrowband system with F = 1 RB, Fig. 1 shows that the proposed method with fixed pilot and cluster allocations (slightly) outperforms the reassignment scheme. The proposed method has the advantage that each UE k is connected to all (at most Q) RUs with the largest LSFCs. Severe pilot contamination is then prevented by the conflict graph. In contrast, a scheduled UE k using the reassignment scheme might end up being connected to only a fraction of possible serving RUs. This can happen since conflicts are avoided by not associating a UE k to a possible RU  $\ell$  if that RU already serves another UE k' with the same pilot and a non-orthogonal channel subspace, i.e.,  $|S_{\ell,k} \cap S_{\ell,k'}| > 0$ . In this way however, severe interference is not prevented.

# B. PFS in a Wideband System

We compare the performance of a wideband system with F = 8 RBs to the narrowband system with F = 1 using the proposed pilot and cluster allocation method. Fig. 2 shows that because of coding over F = 8 RBs in (2), the empirical CDF of the instantaneous mutual information behaves in a more deterministic way. This allows a more aggressive instantaneous rate allocation in the active slots. As a result, see Fig. 2, the user throughput rate in a system with F = 8 can be significantly increased compared to F = 1.

#### C. Concluding Remarks

In this work, we considered a user-centric cell-free massive MIMO network with a total number of users in its area that is much larger than the optimal user load. We proposed a fixed pilot and cluster assignment scheme under scheduling, which greatly reduces the communication overhead between UEs and RUs, while also achieving better performance compared to the scheme in [9]. We further showed that when coding over several RBs in a wideband system, the mutual information can be predicted more accurately, yielding a smaller probability of information outage. This in turn leads to a larger UE throughput compared to the narrowband system with one RB.

## REFERENCES

- G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Trans. on Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.
- [2] 3GPP, "Physical channels and modulation (Release 16)," 3GPP Technical Specification 38.211, 12 2020, Version 16.4.0.
- [3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. on Wireless Comm.*, vol. 9, no. 11, pp. 3590–3600, 2010.
- [4] 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.
- [5] E. Björnson and L. Sanguinetti, "Scalable Cell-Free Massive MIMO Systems," *IEEE Trans. on Comm.*, vol. 68, no. 7, pp. 4247–4261, 2020.
- [6] S. Chen, J. Zhang, E. Björnson, and B. Ai, "Improving Fairness for Cell-Free Massive MIMO Through Interference-Aware Massive Access," *IEEE Transactions on Vehicular Technology*, pp. 1–6, 2022.
- [7] F. Göttsch, N. Osawa, T. Ohseki, K. Yamazaki, and G. Caire, "Subspace-Based Pilot Decontamination in User-Centric Scalable Cell-Free Wireless Networks," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.
- [8] A. Forenza, S. Perlman, F. Saibi, M. Di Dio, R. van der Laan, and G. Caire, "Achieving large multiplexing gain in distributed antenna systems via cooperation with pCell technology," in 2015 49th Asilomar Conference on Signals, Systems and Computers, 2015, pp. 286–293.
- [9] F. Göttsch, N. Osawa, T. Ohseki, Y. Amano, I. Kanno, K. Yamazaki, and G. Caire, "Fairness Scheduling in Dense User-Centric Cell-Free Massive MIMO Networks," arXiv preprint arXiv:2211.15294, 2022.
- [10] H. Shirani-Mehr, G. Caire, and M. J. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *IEEE Transactions on Communications*, vol. 58, no. 7, pp. 2055–2066, 2010.
- [11] Z. Chen, E. Björnson, and E. G. Larsson, "Dynamic resource allocation in co-located and cell-free massive MIMO," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 1, pp. 209–220, 2019.
- [12] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "Distributed resource allocation optimization for user-centric cell-free MIMO networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3099–3115, 2021.
- [13] J. Denis and M. Assaad, "Improving Cell-Free Massive MIMO Networks Performance: A User Scheduling Approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7360–7374, 2021.
- [14] L. Georgiadis, M. J. Neely, L. Tassiulas et al., "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends® in Networking*, vol. 1, no. 1, pp. 1–144, 2006.
- [15] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: the large-scale array regime," *IEEE Trans. on Inform. Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.
- [16] F. Göttsch, N. Osawa, T. Ohseki, K. Yamazaki, and G. Caire, "The Impact of Subspace-Based Pilot Decontamination in User-Centric Scalable Cell-Free Wireless Networks," in 2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2021, pp. 406–410.
- [17] —, "Uplink-Downlink Duality and Precoding Strategies with Partial CSI in Cell-Free Wireless Networks," in 2022 IEEE Wireless Communications and Networking Conference (WCNC), 2022, pp. 1–6.
- [18] L. Miretti, R. L. Cavalcante, E. Björnson, and S. Stańczak, "UL-DL duality for cell-free massive MIMO with per-AP power and information constraints," *arXiv preprint arXiv:2301.06520*, 2023.
- [19] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Informationtheoretic and communications aspects," *IEEE transactions on information theory*, vol. 44, no. 6, pp. 2619–2692, 1998.
- [20] N. Osawa, F. Göttsch, I. Kanno, T. Ohseki, Y. Amano, K. Yamazaki, and G. Caire, "Effective Overloaded Pilot Assignment with Pilot Decontamination for Cell-Free Systems," arXiv preprint arXiv:2207.11478, 2022.
- [21] F. Ye, J. Li, P. Zhu, D. Wang, H. Wu, and X. You, "Spectral Efficiency Analysis of Cell-Free Distributed Massive MIMO Systems With Imperfect Covariance Matrix," *IEEE Systems Journal*, vol. 16, no. 4, pp. 5402–5412, 2022.