Scheduling Strategies for Hierarchical Beamforming in Cloud RAN

Boriana Boiadjieva, Hussein Al-Shatri and Anja Klein
Communications Engineering Lab
Technische Universität Darmstadt, Merckstrasse 25, 64283 Darmstadt, Germany
{b.boiadjieva, h.shatri, a.klein}@nt.tu-darmstadt.de

***Abstract*—In this paper, we consider the downlink of a multicellular multiuser system with cloud radio access network (CRAN) and study the user scheduling with hierarchical beamforming which maximizes the system sum rate. To reduce the signaling over the capacity limited fronthaul, we feed the cloud only with long-term statistical channel information. Therefore, we use theorems from random matrix theory to define tight analytical approximation of the data rate at every user. Hence, we split the preprocessing between cloud and base stations (BSs) to allow intra-cell and inter-cell interference management and apply hierarchical beamforming designed partly at the cloud and partly at the BS. Having this split, we posed the intuitive questions where and how the user scheduling should be applied and study which strategy can provide us with the best trade-off between achieved system sum rate performance and required execution time. Since the optimal solution will require high computational complexity as well as a big amount of signaling over the fronthaul, we investigate diverse suboptimal schemes for user scheduling at the cloud and at the BSs. The proposed schemes have the objective to maximize the system sum rate. Interestingly, although based only on statistical knowledge, simulation results demonstrate that the scheduling at the cloud outperforms scheduling at the BSs since it allows coordination for interference management between the BSs and so it achieves higher system sum rate while keeping the computational time low.**

I. Introduction

Future mobile cellular networks will support a huge number of devices as well as a huge amount of data traffic [1]. To meet the growing demands, the mobile network should take advantage of diverse engineering solutions, many of which will lead to densified cells and application of large antenna arrays at the base stations (BSs) implying multiple input multiple output (MIMO) techniques. On the other hand, the big number of cells, antennas and users in the network will increase the interference and so the complexity of BS coordination. Therefore, the new promising architecture, called cloud radio access network (CRAN), has been proposed to provide centralization and coordination among the BSs through flexible baseband processing and functional splits for fully centralized and hybrid solutions. Besides the advanced coordination, CRAN comes with the challenge related to the big amount of signaling over the fronthaul links, connecting the BSs with the cloud. In practice, the fronthaul is capacity constrained and time-delay constrained [2], therefore the design of channel state information (CSI) based techniques, like the coordinated scheduling/ coordinated beamforming, should be carefully done by taking into account the fronthaul constrains.

To address this challenge, different proposals can be found from the research community suggesting compressive CSI acquisition and fronthaul compression strategies, see [3] and references therein. For example, in [4], we proposed a hierarchical beamforming where the cloud receives only averaged channel link qualities and designs the transmission subspace for every BS. Another example is [5], where a new CSI acquisition scheme is proposed to reduce the CSI overhead and to allow statistical beamforming at the cloud. In [6], we consider the downlink of multicellular system and feed the cloud only with long-term channel statistics which vary significantly slower than the instantaneous CSI. This results in a network in which the BSs have instantaneous but only local CSI, i.e., CSI of their own users, while the cloud has global CSI, i.e., CSI of the whole system, but only statistical channel knowledge. This allows the design of a top-down one-shot hierarchical beamformer, partly designed at the cloud and partly at the BSs, which maximizes the system sum rate. Moreover, since the cloud has only statistical channel knowledge, which is mathematically described through large random matrices, we apply random matrix theory (RMT) to allow analysis and design of the system. This theory provides us with deterministic equivalents which are closed-form expressions, approximating tightly the random processes and thus allowing the cloud to perform accurate prediction and BS coordination.

Since the proposed hierarchical beamforming considers less users than the number of antennas and serves all of them simultaneously, in this paper we apply user scheduling to obtain a more practical preprocessing which is generalized for any number of users. However, applying conventional schemes at the BS which depend on the local CSI might lead to performance degradation due to the lack of inter-cell interference coordination. Additionally, the attention to reduce the CSI overhead and the complexity in the preprocessing has become greater since the number of antennas and number of users increase drastically with the new coming generation. Therefore, many researches study preprocessing techniques based on statistical CSI. In [7], a multi-user single cell MIMO system is considered and based on the Mullen's inequality, a lower bound on the average signal-to-leakage-and-noise ratio has been derived to give an analytical expression of the ergodic sum rate. Having this, the authors propose a downlink transmission where the BSs have only statistical CSI

and separate the users by the so-called statistical-eigenmode space-division multiple-access. In [8], the authors show an effective method to combine long-term statistical CSI with instantaneous low rate feedback and propose a low complexity joint scheduling and beamforming with statistical CSI for the downlink of a single cell. Another example is [9], where joint statistical beamforming and user scheduling is proposed. For the scheduling, an analytical sum rate expression, based on asymptotic approximations on the channel covariance, has been derived. Thus, the statistical CSI is a very promising solution for signaling reduction over the fronthaul while enabling coordinated user scheduling at the cloud. Considering the CSI split between BSs and cloud introduced in our previous work [6], we study the user scheduling for hierarchical beamforming and aim to find out where in the system the scheduling should be applied: at the cloud which has global but only statistical CSI, or at the BSs which have instantaneous but only local CSI within the transmission subspace predefined by the cloud. Aiming at the maximization of the system sum rate, we propose four different scheduling strategies, applied either at the cloud or at the BS, and study their performance by comparing the achieved system sum rate and the required execution time.

The paper is organized as follows. In Section II, we introduce the system model. In Section III, we show the hierarchical beamforming where Subsection III-A presents the deterministic approximations of the power terms at every user in the system and Subsection III-B the outer beamformer design. In Section IV, we describe the proposed scheduling strategies and in Section V, we show simulation results. Section VI concludes the work.

Notations - To denote vectors and matrices, we use lower case and upper case boldface letters, respectively. The ith entry of the vector $\mathbf{x}$ and the (i,j)th entry of the matrix $\mathbf{X}$ are denoted by $[\mathbf{x}]_i$ and $[\mathbf{X}]_{i,j}$, respectively. An $N \times N$ diagonal matrix with entries of $\mathbf{x}$ is denoted by $\mathrm{diag}(\mathbf{x})$. The identity matrix of size $N \times N$ is denoted by $\mathbf{I}_N$. Furthermore, $(\cdot)^{\mathrm{H}}$ stands for the Hermitian of a matrix, $\mathrm{tr}(\cdot)$ for the trace of a matrix and $||\mathbf{x}||$ for the Euclidean norm of vector $\mathbf{x}$. $|\mathcal{A}|$ denotes the cardinality of a set $\mathcal{A}$ and $\mathring{x}$ the deterministic equivalent of a functional x.

II. System model

In this paper, we consider the downlink of a multicellular network with L BSs coordinated by the cloud. Every BS has M_l antennas and simultaneously serves a set $\mathcal{K}_l$ of single-antenna users chosen by the user scheduler from the set $\mathcal{U}_l$ such that $K_l = |\mathcal{K}_l|$ and $K_l \leq M_l$. The serving BS of user k is denoted by $l_k \in \{1, \ldots, L\}$. To user k, it transmits the symbol s_{k,l_k}, modeled as zero mean Gaussian process with variance one, i.e. $s_{k,l_k} \sim \mathcal{CN}(0,1)$. Considering the one-ring channel model [10], we denote the channel vector between BS l and user k by $\mathbf{h}_{k,l} = \sqrt{a_{k,l}}\bar{\mathbf{\Theta}}_{k,l}^{1/2}\mathbf{z}_{k,l} \in \mathbb{C}^{M_l \times 1}$ with $a_{k,l}$ the long-term path loss, $\bar{\mathbf{\Theta}}_{k,l} \in \mathbb{C}^{M_l \times M_l}$ the correlation matrix of the channel and $\mathbf{z}_{k,l} \in \mathbb{C}^{M_l \times 1} \sim \mathcal{CN}(0, \mathbf{I}_{M_l})$ describing the fast channel fluctuations. The long-term statistics of the channel $\mathbf{\Theta}_{k,l} = a_{k,l}\bar{\mathbf{\Theta}}_{k,l} \in \mathbb{C}^{M_l \times M_l}$ vary few orders slower than the fast channel fluctuations $\mathbf{z}_{k,l}$. Additionally, we assume that the users are separated by at least few wavelengths which means that their channels are mutually independent.

The beamforming vector at BS l for user k is denoted by $\mathbf{v}_{k,l} \in \mathbb{C}^{M_l \times 1}$ such that the beamformer at BS l is $\mathbf{V}_l = [\mathbf{v}_{i,l}]_{i \in \mathcal{K}_l} \in \mathbb{C}^{M_l \times K_l}$. The power allocated at BS l for user k is denoted by $p_{k,l}$ and the noise at user k as n_k modeled as zero mean white Gaussian process with variance $\sigma^2 = 1$. Therefore, the resulting received signal at user $k \in \mathcal{K}_l$ consists of four components: useful signal, intra-cell interference, inter-cell interference and noise components, respectively, and it is described by

$$y_k = \mathbf{h}_{k,l_k}^{\mathrm{H}}\sqrt{p_{k,l_k}}\mathbf{v}_{k,l_k}s_{k,l_k} + \sum_{i \in \mathcal{K}_{l_k}, i \neq k} \mathbf{h}_{k,l_k}^{\mathrm{H}}\sqrt{p_{i,l_k}}\mathbf{v}_{i,l_k}s_{i,l_k} + \sum_{\substack{l=1,\\ l \neq l_k}}^{L} \sum_{j \in \mathcal{K}_l} \mathbf{h}_{k,l}^{\mathrm{H}}\sqrt{p_{j,l}}\mathbf{v}_{j,l}s_{j,l} + n_k. \quad (1)$$

The data rate R_k in bit/s/Hz at user $k \in \mathcal{K}_{l_k}$ depends on the received useful power S_k, intra-cell I_k^{ra} and inter-cell I_k^{er} power terms such that

$$S_k = p_{k,l_k}|\mathbf{h}_{k,l_k}^{\mathrm{H}}\mathbf{v}_{k,l_k}|^2, \quad (2)$$

$$I_k^{ra} = \sum_{i \in \mathcal{K}_{l_k}, i \neq k} p_{i,l_k}|\mathbf{h}_{k,l_k}^{\mathrm{H}}\mathbf{v}_{i,l_k}|^2, \quad (3)$$

$$I_k^{er} = \sum_{l=1, l \neq l_k}^{L} \sum_{j \in \mathcal{K}_l} p_{j,l}|\mathbf{h}_{k,l}^{\mathrm{H}}\mathbf{v}_{j,l}|^2, \quad (4)$$

$$R_k = \log_2(1 + S_k/(I_k^{ra} + I_k^{er} + \sigma^2)). \quad (5)$$

III. Hierarchical Beamforming

In this section, the hierarchical beamforming is explained. In [6], we have proposed this beamforming for the scenario of multicellular multiuser CRAN system where the number of the users in each cell does not exceed the number of antennas at the BS and, hence, all users are served simultaneously without the need of user scheduling. In this section, we summarize the main hierarchical beamforming concept as far as needed to perform the new preprocessing design proposed in this paper, namely user scheduling for hierarchical beamforming.

The hierarchical beamforming consists of two concatenated beamformers, i.e. $\mathbf{V}_l = \mathbf{F}_l\mathbf{G}_l$ with $\mathbf{F}_l \in \mathbb{C}^{M_l \times M_l}$ and $\mathbf{G}_l = [\mathbf{g}_{i,l}]_{i \in \mathcal{K}_l} \in \mathbb{C}^{M_l \times K_l}$. The so-called outer beamformer $\mathbf{F}_l$ is designed at the cloud together with all other outer beamformers in order to allow coordination. It is based only on the available statistics at the cloud and it defines the transmission subspace for the corresponding BSs. The second, so-called inner beamformer $\mathbf{G}_l$ is designed locally at the BSs based only on its local but instantaneous CSI.

In order to keep the complexity and the fronthaul transmissions as low as possible, in [6], we have designed a top-down one-shot hierarchical beamformer where the inner beamformer is of closed-form and the outer beamformer is designed to maximize the system sum rate. More precisely, the inner

beamformer applies the regularized zero-forcing (RZF) [11], [12] with $\xi_l^2 = P_l/\mathrm{tr}(\mathbf{P}_l\bar{\mathbf{G}}_l^{\mathrm{H}}\bar{\mathbf{G}}_l)$ a normalization parameter applied to fulfill the power budget P_l at BS l and $\bar{\mathbf{G}}_l$ the non-normalized inner beamforming. $\mathbf{P}_l = \mathrm{diag}(\mathbf{p}_l) \in \mathbb{R}_+^{K_l \times K_l}$ is the power allocation matrix at BS l with $\mathbf{p}_l = [p_{i,l}]_{i\in\mathcal{K}_l} \in \mathbb{C}^{K_l\times 1}$. For this beamforming, we use the effective channel matrix $\tilde{\mathbf{H}}_l = [\tilde{\mathbf{h}}_{i,l}]_{i\in\mathcal{K}_l}^{\mathrm{H}} \in \mathbb{C}^{K_l\times M_l}$ where $\tilde{\mathbf{h}}_{k,l} = \mathbf{F}_l^{\mathrm{H}}\mathbf{h}_{k,l}$ is the effective channel within the transmission subspace between user k and BS l. RZF uses a regularization parameter α_l which controls the interference in the cell and it has been chosen to maximize the signal to interference and noise ratio in single cell scenario with only local channel knowledge [13], [14]: $\alpha_l = (K_l\sigma^2)/(P_lM_l)$. Hence, the resulting inner beamformer is described by

$$\mathbf{G}_l = \xi_l\bar{\mathbf{G}}_l = \xi_l\big(\tilde{\mathbf{H}}_l^{\mathrm{H}}\tilde{\mathbf{H}}_l + M_l\alpha_l\mathbf{I}_{M_l}\big)^{-1}\tilde{\mathbf{H}}_l^{\mathrm{H}}. \tag{6}$$

Since the inner beamformer is of closed-form, the cloud can design the outer beamformers without the need of knowledge of any inner beamformer realization as well as without the knowledge of the instantaneous CSI. To achieve this, in [6], we apply RMT and more precisely the method of the deterministic equivalents which provides us with closed-form expressions approximating the data rate at every user in the system based only on the statistical knowledge $\mathbf{\Theta}_{k,l}$ for $\forall k, l$.

A. Deterministic Equivalents

Since the radio channel $\mathbf{H}_l = [\mathbf{h}_{i,l}]_{i\in\mathcal{K}_l}$ of BS l is a random matrix of size $K_l \times M_l$, we use random matrix theory to obtain deterministic equivalents of the data rate at every user. The deterministic equivalents are asymptotic expressions of functionals with random matrices whose dimensions approach infinity [15]. These expressions are of closed-form and approximate tightly the random processes for systems of finite size and even for those of very small dimensions. For our system, the deterministic equivalent of a functional x, which depends on the radio channels $\mathbf{H}_l$, is denoted as $\mathring{x}$, which depends only on the channels' second order statistics $\mathbf{\Theta}_{k,l}$ such that $x - \mathring{x} \xrightarrow{a.s.} 0$. The notation "$\xrightarrow{a.s.}$" refers to almost sure convergence in the limit $M_l, K_l \to \infty$ such that $0 < \liminf_{M_l,K_l}\beta_l \le \limsup_{M_l,K_l}\beta_l < \infty$ for $l = 1,\ldots,L$ and $\beta_l = M_l/K_l$ the cell loading.

Through RMT lemmas and the continuous mapping theorem [16], in [6], we have derived the deterministic equivalent of the data rate R_k at any user k in the multicellular system and the deterministic equivalent of the overall system sum rate $R_{\mathrm{sum}} = \sum_{l=1}^{L}\sum_{k\in\mathcal{K}_l} R_k$.

Therefore, the deterministic equivalent of the data rate at user k is defined as $\mathring{R}_k = \log_2(1 + \mathring{S}_k/(\mathring{I}_k^{ra} + \mathring{I}_k^{er} + \sigma^2))$ such that $R_k - \mathring{R}_k \xrightarrow{a.s.} 0$ and the deterministic equivalent of the overall system sum rate as $\mathring{R}_{\mathrm{sum}} = \sum_{l=1}^{L}\sum_{k\in\mathcal{K}_l}\mathring{R}_k$ such that $R_{\mathrm{sum}} - \mathring{R}_{\mathrm{sum}} \xrightarrow{a.s.} 0$. Based on these results, the cloud is able to analyze the system, to design the outer beamformers and to schedule the users, without the actual need of any instantaneous channel or inner beamformer realization.

In order to define the deterministic equivalents of the equations (2) - (5), first we need to calculate the following set of equations:

$$\mathbf{e}_l = \left[\frac{1}{M_l}\mathrm{tr}(\mathbf{F}_l^{\mathrm{H}}\mathbf{\Theta}_{i,l}\mathbf{F}_l\mathbf{T}_l)\right]_{i\in\mathcal{K}_l} \in \mathbb{C}^{K_l\times 1}, \tag{7a}$$

$$\mathbf{T}_l = \left(\frac{1}{M_l}\sum_{j\in\mathcal{K}_l}\frac{\mathbf{F}_l^{\mathrm{H}}\mathbf{\Theta}_{j,l}\mathbf{F}_l}{(1+[\mathbf{e}_l]_j)} + \alpha_l\mathbf{I}_{M_l}\right)^{-1}, \tag{7b}$$

$$\mathring{\Psi}_l = \frac{1}{M_l}\sum_{j\in\mathcal{K}_l}\frac{p_{j,l}[\mathbf{e}_l']_j}{(1+[\mathbf{e}_l]_j)^2}, \tag{7c}$$

$$\mathbf{e}_l' = \mathbf{D}_l\mathbf{v}_l, \tag{7d}$$

$$\mathbf{D}_l = (\mathbf{I}_{K_l} - \mathbf{J}_l)^{-1}, \tag{7e}$$

$$\mathbf{v}_l = \left[\frac{1}{M_l}\mathrm{tr}(\mathbf{F}_l^{\mathrm{H}}\mathbf{\Theta}_{t,l}\mathbf{F}_l\mathbf{T}_l^2)\right]_{t\in\mathcal{K}_l} \in \mathbb{C}^{K_l\times 1}, \tag{7f}$$

$$[\mathbf{J}_l]_{i,j} = \frac{\mathrm{tr}(\mathbf{F}_l^{\mathrm{H}}\mathbf{\Theta}_{i,l}\mathbf{F}_l\mathbf{T}_l\mathbf{F}_l^{\mathrm{H}}\mathbf{\Theta}_{j,l}\mathbf{F}_l\mathbf{T}_l)}{M_l^2(1+[\mathbf{e}_l]_j)^2} \text{ for } i,j\in\mathcal{K}_l, \tag{7g}$$

$$\mathring{\Upsilon}_{k,l} = \begin{cases} \sum_{j\in\mathcal{K}_l, j\neq k}\frac{p_{j,l}[\mathbf{c}_{k,l}']_j}{(1+[\mathbf{e}_l]_j)^2} & \text{for } l = l_k \\ \sum_{j\in\mathcal{K}_l}\frac{p_{j,l}[\mathbf{c}_{k,l}']_j}{(1+[\mathbf{e}_l]_j)^2} & \text{otherwise} \end{cases}, \tag{7h}$$

$$\mathbf{c}_{k,l}' = \mathbf{D}_l\mathbf{w}_{k,l}, \tag{7i}$$

$$\mathbf{w}_{k,l} = \left[\frac{1}{M_l}\mathrm{tr}(\mathbf{F}_l^{\mathrm{H}}\mathbf{\Theta}_{t,l}\mathbf{F}_l\mathbf{T}_l\mathbf{F}_l^{\mathrm{H}}\mathbf{\Theta}_{k,l}\mathbf{F}_l\mathbf{T}_l)\right]_{t\in\mathcal{K}_l} \in \mathbb{C}^{K_l\times 1}. \tag{7j}$$

Having defined the terms from (7), we can obtain the deterministic equivalents of all the power terms at user k as :

$$\mathring{S}_k = \frac{p_{k,l_k}P_{l_k}[\mathbf{e}_{l_k}]_k^2}{\mathring{\Psi}_{l_k}(1+[\mathbf{e}_{l_k}]_k)^2}, \tag{8}$$

$$\mathring{I}_k^{ra} = \frac{P_{l_k}\mathring{\Upsilon}_{k,l_k}}{M_{l_k}\mathring{\Psi}_{l_k}(1+[\mathbf{e}_{l_k}]_k)^2}, \tag{9}$$

$$\mathring{I}_k^{er} = \sum_{l=1,l\neq l_k}^{L}\frac{P_l}{M_l\mathring{\Psi}_l}\mathring{\Upsilon}_{k,l}. \tag{10}$$

Proof: see [6].

B. Outer Beamformer Design

The outer beamformers are designed so that they maximize the system sum rate $\mathring{R}_{\mathrm{sum}}$ following a low complexity algorithm based on interactive block diagonalization, see [6]. The algorithm searches iteratively over all BSs for subspaces which are orthogonal or nearly orthogonal to the interference producing subspace of a BS b, i.e. $\mathbf{B}_b^i = \left[\mathbf{F}_b^{\mathrm{H}}\mathbf{\Theta}_{j,b}\mathbf{F}_b\right]_{j\in\{\mathcal{K}_l: l=1,\ldots,L \text{ and } l\neq b\}}$, and additionally it considers only the strongest modes from the serving subspace $\mathbf{B}_b^s = \left[\mathbf{F}_b^{\mathrm{H}}\mathbf{\Theta}_{i,b}\mathbf{F}_b\right]_{i\in\mathcal{K}_b}$ while in the same time it takes into account all the subspaces from the previous iterations. The search for best subspaces terminates when the cloud cannot find a better set of transmission subspaces leading to higher system sum rate.

In one iteration, the first step is to perform singular value decomposition of the matrix $\mathbf{B}_b^i$ and to denote the left singular vectors as $\mathbf{E}_b$. Then, we define $\mathbf{E}_b^0$ to be a matrix which

collects the vectors $\mathbf{E}_b$ corresponding to the weakest N_b^i singular values. Afterwards, we project the subspace $\mathbf{B}_b^s$ onto $\mathbf{E}_b^0$ and denote the projection by $\mathbf{M}_b$. As a next step, we take only the strongest N_b^s eigenmodes which define the matrix $\mathbf{M}_b^1$. Hence, the outer beamformer at BS b is designed to be $\mathbf{F}_b = \mathbf{E}_b^0\mathbf{M}_b^1$. N_b^s and N_b^i are chosen at each iteration so that the resulting sum rate is maximized.

IV. Scheduling Strategies

Our main goal is to achieve maximum system sum rate R_{sum} with respect to the sets $\mathcal{K}_1 \subseteq \mathcal{U}_1, \ldots, \mathcal{K}_L \subseteq \mathcal{U}_L$ of served users, the outer beamformers $\mathbf{F}_1, \ldots, \mathbf{F}_L$ and the inner beamformers $\mathbf{G}_1, \ldots, \mathbf{G}_L$ which is in general computationally intractable. Therefore, we propose four sub-optimal algorithms which differ by user scheduling, applied either at the cloud or at the BSs and do not require additional signaling, i.e. the same channel information which is used for beamforming is exploited for user scheduling. We investigate where should the user scheduling be applied and which algorithm can provide us with the best performance considering the achieved system sum rate and the required execution time.

The algorithms are described in the subsections below and their main principles are briefly captured in Fig. 1. Three of the scheduling algorithms are iterative, therefore we use the index i to represent the iteration index. In this section, we annotate the deterministic equivalent of the sum rate as a function of the served users sets, i.e. $\mathring{R}_{\text{sum}}(\{\mathcal{K}_{1,i}, \ldots, \mathcal{K}_{L,i}\}) = \sum_{l=1}^{L}\sum_{k\in\mathcal{K}_l}\mathring{R}_k$ to emphasize on which user sets the data rate depends. We also annotate the data rate $\mathring{R}_k$ as $\mathring{R}_k\big(\mathbf{F}_{1,i}, \ldots, \mathbf{F}_{L,i}\big)$ for the strategy in which the outer beamformers are calculated in every iteration i.

For the algorithm with scheduling at the BS, the data rate in cell l is $R_l^{\text{cell}} = \sum_{k\in\mathcal{K}_l}\log(1+\tilde{S}_k/(I_k^{\tilde{r}a}+\sigma^2)) = \sum_{k\in\mathcal{K}_l}R_k^{\text{BS}}$ where $\tilde{S}_k = p_{k,l_k}|\tilde{\mathbf{h}}_{k,l_k}^{\text{H}}\mathbf{g}_{k,l_k}|^2$ and $I_k^{\tilde{r}a} = \sum_{i\in\mathcal{K}_{l_k}, i\neq k}^{L} p_{i,l_k}|\tilde{\mathbf{h}}_{k,l_k}^{\text{H}}\mathbf{g}_{i,l_k}|^2$. In the desiction metric we have $R_k^{\text{BS}}(\mathbf{G}_{l,i})$ which is the data rate defined at the BS for user k which does not account for the received inter-cell interference since the BS does not have this information and it depends on the inner beamformer $\mathbf{G}_{l,i}$ calculated in each iteration.

Additionally, we introduce the set $\mathcal{B}$ which is a set used from the iterative algorithms with user scheduling at the cloud and which shows the BSs whose user sets have to be updated. The maximum number of served users at BS l is $K_l^{\max} = \min\{M_l, |\mathcal{U}_l|\}$.

A. Scheduling at the cloud (SC)

For SC, see Fig. 1, the user scheduler is the first step of the preprocessing at the cloud and it assigns the users in the system iteratively over all BSs which need to be updated, i.e. these which are in the set $\mathcal{B}$. Because the preprocessing starts with scheduling, there are no designed outer beamformers and the scheduler does not consider their realization, i.e. $\mathbf{F}_l = \mathbf{I}_{M_l}$ for $\forall l$. Therefore, in Fig. 1, it is emphasized that $\mathring{R}_{\text{sum}}$ depends only on the user sets. In every iteration, the user set $\mathcal{K}_b$ at BS $b \in \mathcal{B}$ is optimized by taking into account the user selections $\mathcal{K}_l$ for $l = 1, \ldots, L$ and $l \neq b$ from the previous iterations. The algorithm updates consecutively every BS's serving set by adding a new user in the system and all BSs repeatedly. The algorithm will stop updating the user set of a BS b when the update of $\mathcal{K}_b$ leads to a deterioration in the system sum rate, here measured by the sum rate difference $\Delta\mathring{R}_{\text{sum},i}$, or when the maximum number of served users $K_b^{\max}$ is achieved.

SC: scheduling at the cloud

1: **initialize**

$$\mathcal{B} = \{1, \ldots, L\}, \mathcal{K}_{l,i=0} = \emptyset, \mathring{R}_{\text{sum},i=0} = 0, i = 1$$

while $i \leq i_{\max}$ and $\mathcal{B} \neq \emptyset$

2: set BS $b \in \mathcal{B}$ to be updated

3: find a user s_i such that

$$s_i = \underset{u\in\mathcal{U}_b\setminus\mathcal{K}_{b,i-1}}{\arg\max}\ \mathring{R}_{\text{sum}}(\{\mathcal{K}_{1,i-1}, \ldots, \mathcal{K}_{L,i-1}\}\cup\{u\})$$

4: compute the decision metric

$$\begin{aligned}\mathring{R}_{\text{sum},i} &= \sum_{l=1}^{L}\left(\sum_{k\in\mathcal{S}_i}\mathring{R}_k\right)\\ \Delta\mathring{R}_{\text{sum},i} &= \mathring{R}_{\text{sum},i} - \mathring{R}_{\text{sum},i-1}\end{aligned}$$

for $\mathcal{S}_i = \{\mathcal{K}_{1,i-1}, \ldots, \mathcal{K}_{L,i-1}\}\cup\{s_i\}$.

5: check whether BS b should be further updated

if $\Delta\mathring{R}_{\text{sum},i} \geq 0$ **then** $\mathcal{K}_{b,i} = \mathcal{K}_{b,i-1}\cup\{s_i\}$

else $\mathcal{K}_{b,i} = \mathcal{K}_{b,i-1}$ and $\mathcal{B} = \mathcal{B}\setminus\{b\}$

if $|\mathcal{K}_{b,i}| = K_b^{\max}$ **then** $\mathcal{B} = \mathcal{B}\setminus\{b\}$

end

After all user sets have been optimized, the cloud designs the outer beamformers considering the selected serving sets. As a next step, the cloud transmits the user sets and the outer beamformers to the corresponding BSs. At BS l, the selected users $\mathcal{K}_l$ and the outer beamformer $\mathbf{F}_l$ are used for the whole time duration in which the statistical CSI does not change. Within this frame of constant channel statistics, the BS designs inner beamformer $\mathbf{G}_l$ for every change in the instantaneous CSI of the effective channel within the transmission subspace defined by the cloud.

B. Scheduling at the cloud with outer beamforers $\underline{\mathbf{F}}_l$ (SC-F)

SC-F updates the user sets iteratively over all BSs in $\mathcal{B}$ analogically to the previous algorithm SC. However, in SC-F, in the decision metric for the user scheduling the calculation of the sum rate difference $\Delta\mathring{R}_{\text{sum},i}$ accounts for the outer beamformers, and hence, their realizations are also computed in every single iteration, see Fig. 1. In the Figure, $\Delta\mathring{R}_{\text{sum},i}$ is expressed as a function of the user sets and outer beamformers in order to emphasize that unlike the other algorithms this one considers not only the user sets but also the outer beamformers. The termination condition for updating a user set is the same as in the SC, i.e., when additional users decrease the sum rate or when the maximum

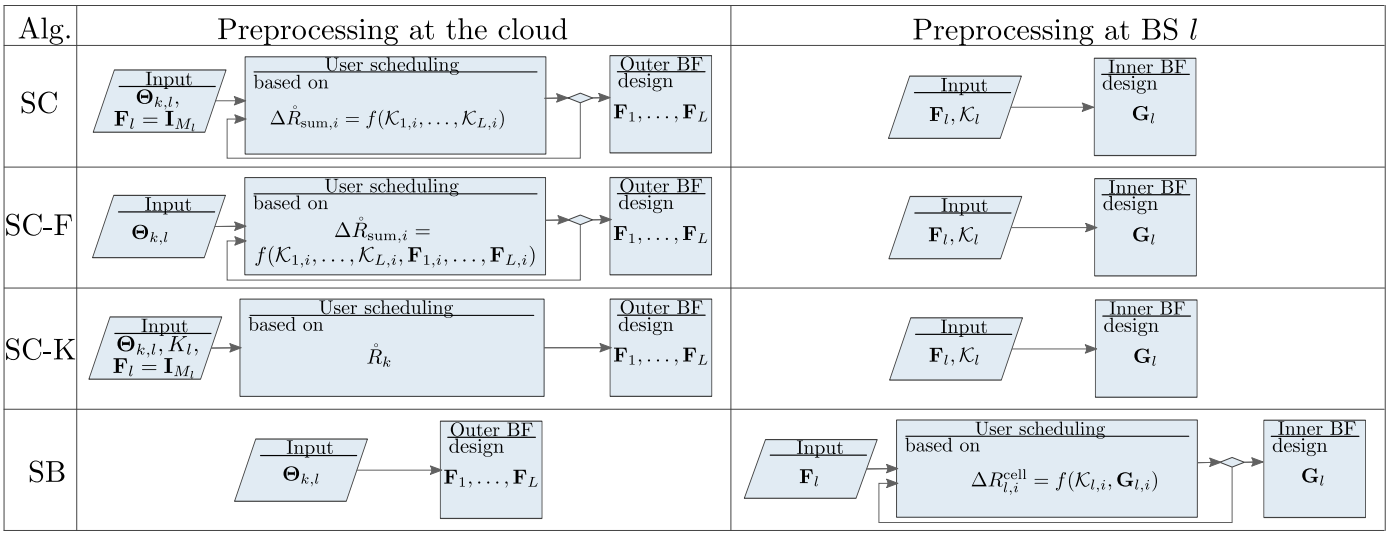

Fig. 1: User scheduling strategies - overview

number of users to be served is achieved. After terminating the user scheduling, the final outer beamformers are designed matched to the best found user selections. The user sets and outer beamformers are then transmitted to the BS for every frame in which the long-term statistical properties remain constant. At every BS, the received outer beamformer and user set are used for the whole frame of constant statistics while the inner beamforming is designed for every single effective channel realization in order to adapt to the fast channel fluctuations within the predefined transmission subspace.

SC-F: scheduling at the cloud

1: **initialize**

$\mathcal{B} = \{1, \ldots, L\}, \mathcal{K}_{l,i=0} = \emptyset, \mathring{R}_{\mathrm{sum},i=0} = 0, \mathbf{F}_{l,i=0} = \mathbf{I}_{M_l}, i = 1$

while $i \leq i_{\max}$ and $\mathcal{B} \neq \emptyset$

2: set BS $b \in \mathcal{B}$ to be updated

3: find a user s_i such that

$$s_i = \underset{u \in \mathcal{U}_b \setminus \mathcal{K}_{b,i-1}}{\arg\max} \; g$$

$g = \mathring{R}_{\mathrm{sum}}(\{\mathcal{K}_{1,i-1}, \ldots, \mathcal{K}_{L,i-1}\} \cup \{u\}, \mathbf{F}_{1,i-1}, \ldots, \mathbf{F}_{L,i-1})$

4: compute the decision metric

$$\mathring{R}_{\mathrm{sum},i} = \sum_{l=1}^{L} \left(\sum_{k \in \mathcal{S}_i} \mathring{R}_k(\mathbf{F}_{1,i}, \ldots, \mathbf{F}_{L,i}) \right)$$

$$\Delta\mathring{R}_{\mathrm{sum},i} = \mathring{R}_{\mathrm{sum},i} - \mathring{R}_{\mathrm{sum},i-1}$$

for $\mathcal{S}_i = \{\mathcal{K}_{1,i-1}, \ldots, \mathcal{K}_{L,i-1}\} \cup \{s_i\}$.

5: check whether BS b should be further updated

if $\Delta\mathring{R}_{\mathrm{sum},i} \geq 0$ **then** $\mathcal{K}_{b,i} = \mathcal{K}_{b,i-1} \cup \{s_i\}$

else $\mathcal{K}_{b,i} = \mathcal{K}_{b,i-1}$ and $\mathcal{B} = \mathcal{B} \setminus \{b\}$

if $|\mathcal{K}_{b,i}| = K_b^{\max}$ **then** $\mathcal{B} = \mathcal{B} \setminus \{b\}$

end

C. Scheduling at the cloud with fixed K_l (SC-K)

In SC-K, the cloud has a predefined number of users to be served for every BS, i.e. $K_1, \ldots, K_L$, which means that the cell loading $\beta_l = M_l/K_l$ at each BS is set to be a constant value, see Fig. 1. Therefore, the cloud schedules only those users who achieve maximum data rate and does not consider the outer beamformers, i.e. $\mathbf{F}_l = \mathbf{I}_{M_l}$. This scheme is not iterative and, therefore, it requires significantly less computational time than the other schemes. After the user scheduling, the outer beamformers are designed at the cloud. Analogical to the previous two algorithms, the user sets and the outer beamformers are transmitted to the BSs in every frame of constant statistical CSI while every BS designs it own inner beamformer based on the instantaneous local CSI of the effective channels.

SC-K: scheduling at the cloud

1: find the K_l users in cell l which give the highest data rate

$$\mathcal{K}_l = \{k \in \mathcal{U}_l : |\{p \in \mathcal{U}_l : \mathring{R}_k < \mathring{R}_p\}| < K_l\}$$

D. Scheduling at the BSs (SB)

In SB, first the outer beamformers are designed at the cloud considering all users in the system and forwarded to the corresponding BS. After that, every BS finds the best user set which maximizes the cell sum rate using the effective instantaneous CSI of its own users $\tilde{\mathbf{h}}_{k,l}$ for $k \in \mathcal{U}_l$. Note that the BSs do not have global channel information, therefore they cannot account for the inter-cell interference caused to the neighboring cells. In Fig. 1, SB is also briefly captured and the dependence of $\Delta R_{l,i}^{\mathrm{cell}}$ on $\mathcal{K}_{l,i}$ and $\mathbf{G}_{l,i}$ is clearly shown.

For all scheduling algorithms at the cloud, we do not have any additional signaling, therefore, the decisions are

taken based only on statistics and using the deterministic equivalents. Additionally, since the statistical CSI, which is the only one available at the cloud, changes very slowly over time, the scheduling at the cloud changes also much slower as compared to the scheduling at the BS where the new user set depends on the instantaneous channel variations.

SB: scheduling at BS l

2: **initialize**

$$\mathcal{K}_{l,i=0} = \emptyset, R^{\text{cell}}_{l,i=0} = 0, i = 1, p = 1$$

while $i \leq i_{\max}$ and $p = 1$

3: find a user s_i such that

$$s_i = \underset{u \in \mathcal{U}_l \setminus \mathcal{K}_{l,i-1}}{\arg\max} \; R^{\text{cell}}_l(\{\mathcal{K}_{l,i-1}\} \cup \{u\}, \mathbf{G}_{l,i})$$

4: compute the decision metric

$$R^{\text{cell}}_{l,i} = \sum_{k \in \{\mathcal{K}_{l,i-1}\} \cup \{s_i\}} R^{\text{BS}}_k(\mathbf{G}_{l,i})$$
$$\Delta R^{\text{cell}}_{l,i} = R^{\text{cell}}_{l,i} - R^{\text{cell}}_{l,i-1}$$

5: check whether BS l should be further updated

if $\Delta R^{\text{cell}}_{l,i} \geq 0$ **then** $\mathcal{K}_{l,i} = \mathcal{K}_{l,i-1} \cup \{s_i\}$

else $\mathcal{K}_{l,i} = \mathcal{K}_{l,i-1}$ and $p = 0$

if $|\mathcal{K}_{l,i}| = K^{\max}_l$ **then** $p = 0$

end

V. Simulation Results

In order to evaluate the performance of the proposed strategies, we execute numerical simulations and compare the achieved system sum rate and the required execution time. Moreover, to obtain a deeper understanding of the algorithms, we examine them under different system conditions such as changing number of available users and different signal to noise ratio (SNR), i.e. power budget over noise power.

A. General Setup

We consider a system which consists of three hexagonal cells each of which with radius $r_{\text{cell}} = 50$ m. In every cell, the users are randomly located following the uniform distribution. We model the path loss between user k and BS l as $a_{k,l} = (d_{k,l}/d_0)^{-\alpha_{loss}}$ where $d_{k,l}$ is the distance between them, d_0 is a reference distance, chosen to be equal to 10 m and $\alpha_{loss} = 3$ describes the path loss exponent. Every BS is placed in the center of one hexagon and distributes its available power equally among the users, i.e. $\mathbf{P}_l = (P_l/K_l)\mathbf{I}_{K_l}$.

The channel correlation $\bar{\mathbf{\Theta}}_{k,l}$ is modeled, using a discrete uniform distribution [17] where $N_{k,l}$ scatterers surround user k with angle of arrival $\phi_{k,l}$ and have angular spread $\Delta_{k,l}$ from the lth BS perspective. Every BS transmits through a uniform linear array with antenna spacing $d = 0.5\lambda$ where λ stand for the carrier wavelength. Therefore, the correlation between the mth and $n$the antenna elements is defined by $[\bar{\mathbf{\Theta}}_{k,l}]_{m,n} = \frac{1}{N_{k,l}} \sum_{i=1}^{N_{k,l}} e^{-j2\pi \frac{d}{\lambda}(m-n)\cos(\theta_{k,l,i})}$ with $\theta_{k,l,i}$ the angle of arrival of the ith scatterer of user k with respect to BS l. To simulate a more realistic system scenario, we let the scattering environment change from frame to frame and model the angular spread as a random variable which is uniformly distributed between 1 and 60 degree, i.e. $\Delta_{k,l} \sim \mathcal{U}(1°, 60°)$, as well as number of scatterers changing with the angular spread, i.e. $N_{k,l} = \Delta_{k,l}$.

In order to achieve an average performance, we execute all simulations for 200 frames where within one frame, the long-term channel statistical properties remain constant. Moreover, every frame consists of 200 random channel realizations. For the iterative algorithms, we set the maximum number of iterations to be $i_{\max} = 100$.

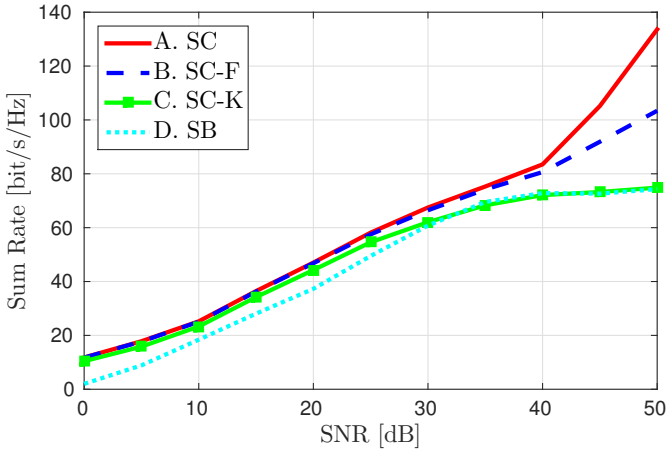

Fig. 2: Average sum rate for $M_l = 8$, $|\mathcal{U}_l| = 16$

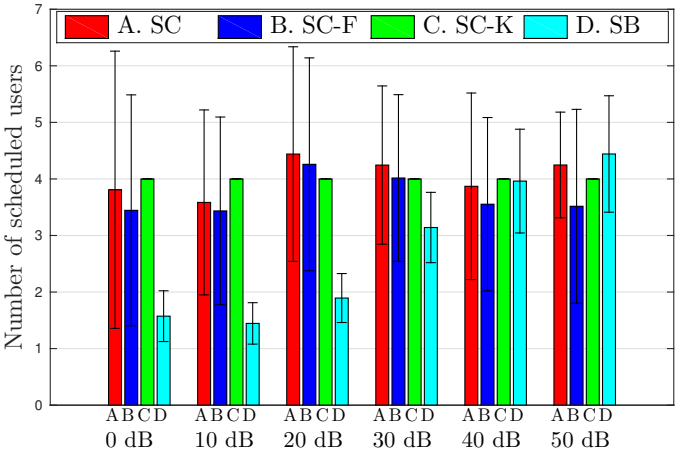

Fig. 3: Average number of served users for $M_l = 8$, $|\mathcal{U}_l| = 16$

B. Average Performance

Fig. 2 shows the achieved average system sum rate as a function of the SNR for all algorithms for a system with $M_l = 8$ antennas at each BS and $|\mathcal{U}_l| = 16$ available users per cell. SC-K has $K_l = 4$ served users at each BS. As for all simulations, the performance is averaged over 200 frames

with constant statistical CSI and 200 instantaneous channel realizations where in every realization the number of scatterers around every user and the angular spread of scatterers is random variable with uniform distribution. Interestingly, even though the cloud has only statistical channel knowledge and the system has only three BSs which might interfere each other, the scheduling at the cloud is clearly more beneficial. Additionally, the SC-K and SB have similar performance at high SNR because these two strategies have fixed transmission subspace dimensionality, i.e. for SC-K, the scheduler should always assign the best $K_l = 4$ users while in SB, every BS is restricted by its transmission subspace predefined at the cloud. This is in contrast to the transmission subspaces of SC and SC-F which are not fixed and the scheduler might not assign any users in a cell for certain time frames. Another interesting observation is that SC-F which has decision metric considering not only the user sets but also the outer beamformers does not achieve the highest system sum rate. This can be explained by the fact that in each iteration the outer beamformers are calculated and so in the next iteration the decision for updated user set is taken based on the already restricted transmission subspaces from the previous iteration.

In Fig. 3, the average number of assigned users per cell is depicted as well as their standard deviation for the same system scenario with $M_l = 8$ and $|\mathcal{U}_l| = 16$. Note that, the bar graph shows the mean of the number of served users for every algorithm and the vertical lines are not an error in the measured mean values but the standard deviation centered around the mean value of the number of served users over all 200 frames and 200 realizations as well as per BS. In SB, we observe the general trend that the number of assigned users increases with the increase of the SNR due to the increased available power which allows the BS to serve more users. On the other hand, the average number of served users from the SC and SC-F varies over the SNR scale with relatively high variance. These results show that due to the unlimited dimensionality of the transmission subspace for these two algorithms, the scheduling at the cloud adapts better to the current statistical channel conditions and so the system achieves a higher sum rate performance.

In Fig. 4, we examine the average required time for preprocessing, more precisely for outer beamforming design and for user scheduling from the same system scenario as in Fig. 2. The algorithms have not been designed for multicore computing, therefore the executions and time measurements are for single core processing. The time consumption for the design of the inner beamformers has not been added because the inner beamformer is of closed-form, requiring orders of magnitude less time and therefore it can be neglected as compared to the overall preprocessing time. As expected, SC-K has the shortest execution time since it has a predefined number of users to be served and omits iterative user search. This simplifies the computations gradually by simple search for the best K_l users at each BS. Comparing SC with SC-F, we observe a big difference in their execution time due to the consideration of the outer beamformers at each iteration in the

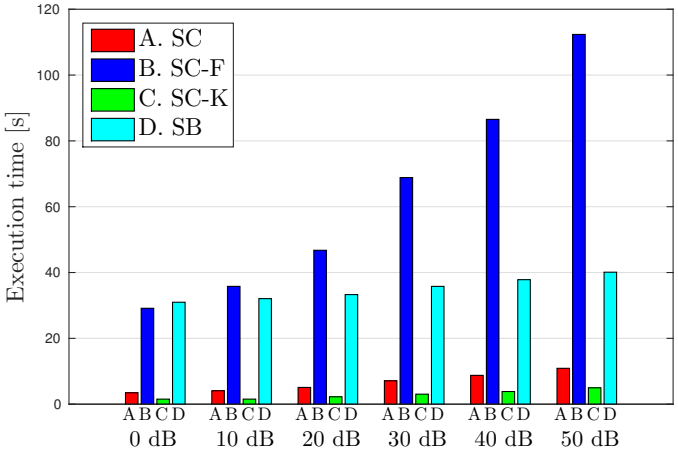

Fig. 4: Average time required for execution for $M_l = 8$, $|\mathcal{U}_l| = 16$

user scheduling of SC-F. The required time for preprocessing from the SB stays significantly high over the whole SNR range. Since the SB algorithm starts its preprocessing with outer beamforming design, the transmission subspaces are designed by considering all users in the system, i.e. $\mathcal{U}_1, \ldots, \mathcal{U}_L$. This leads to very big time investment in designing the outer beamformers because in general, its execution time increases with the increase of the number of antennas and number of users in the system. Comparing the time for user scheduling at the BS, which is multiple times less than the time for outer beamforming design, the main fraction of the overall preprocessing time is due to designing outer beamformers by considering all available users. Note that in each channel realization the maximum time consumption at a BS is taken and not just an average over all BSs' required time. From these results, we can conclude that the outer beamformer design with iterative block diagonalization might introduce big computational burden which can be improved for example by finding the optimal number N_l^s of dimensions to be occupied and not letting the cloud search for the best N_l^s dimensions while designing the subspaces.

In Fig. 5 and Fig. 6, we have a system with SNR $= 25$ dB, three BSs each with $M_l = 8$ antennas and varying number of available users $|\mathcal{U}_l| = 2, \ldots, 64$. Here, SC-K has $K_l = 5$ served users. The results show that the more available users we have in the system, the higher system sum rate can be achieved because the probability to serve good conditioned users, i.e. achieving high data rate by introducing only small portions of interference, is higher. Moreover, in Fig. 6, we observe the trend that the number of served users increases with increasing number of available users. Additionally, for SC and SC-F, the number of served users is roughly half of the number of antennas. In SB, the number of served users is even less because the transmission subspaces defined at the cloud consider all users in the system and as a result restrict the transmission subspace dimensionality a lot, letting only a small number of available dimensions for transmission.

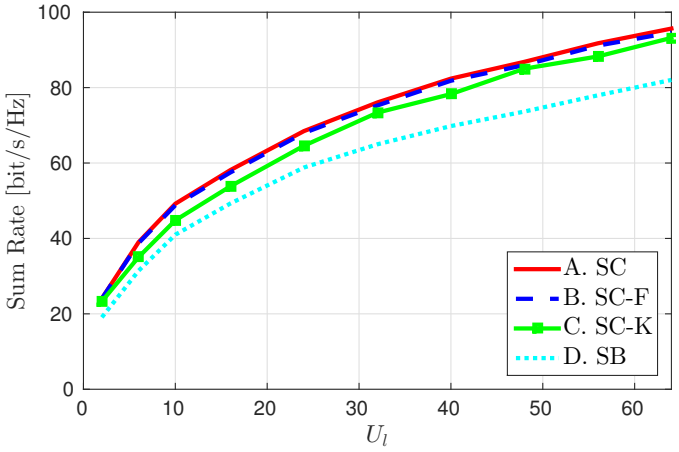

Fig. 5: Average sum rate for SNR 25 dB and $M_l = 8$

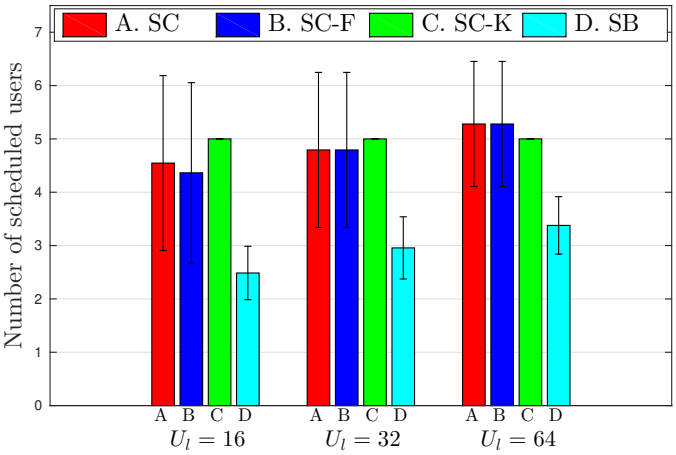

Fig. 6: Average number of scheduled users for SNR 25 dB and $M_l = 8$

VI. Conclusions

In this paper, we considered a system with hierarchical beamforming which is designed partly at the cloud and partly at the BS. This beamforming split allows coordination while requiring only small amount of signaling over the fronthaul links because only statistical CSI is transmitted to the cloud. To generalize and enhance the preprocessing, we study the user scheduling and propose four different scheduling strategies which do not require additional signaling. All algorithms have inner beamformers designed at the BS which adapts to the instantaneous channel changes within the transmission subspace predefined at the cloud. Three of the strategies (SC, SC-F and SC-K) perform coordinated scheduling/ coordinated outer beamforming at the cloud and the fourth one (SB) has coordinated outer beamforming at the cloud but assigns the served users at the BSs. Simulation results show that letting the cloud schedule the users is more beneficial, even though it has only slow varying statistical channel knowledge. The scheduling at the cloud achieves impressive system sum rate while not demanding long executing time.

Acknowledgment

This work has been performed in the context of DFG funded CRC 1053 MAKI.

References

[1] P. Marsch, B. Raaf, A. Szufarska, P. Mogensen, H. Guan, M. Farber, S. Redana, K. Pedersen, and T. Kolding, "Future mobile communication networks: Challenges in the design and operation," *IEEE Vehicular Technology Magazine*, vol. 7, no. 1, pp. 16–23, 2012.

[2] M. Peng, C. Wang, V. Lau, and H. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Communications*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[3] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2282–2308, 2016.

[4] H. Al-Shatri and A. Klein, "Hierarchical beamforming for downlink cloud radio access networks," *14th International Symposium on Wireless Communication Systems*, 2017.

[5] Y. Shi, J. Zhang, and K. B. Letaief, "CSI overhead reduction with stochastic beamforming for cloud radio access networks," *IEEE International Conference on Communications*, pp. 5154–5159, 2014.

[6] B. Boiadjieva, H. Al-Shatri, and A. Klein, "Hierarchical beamforming in CRAN using random matrix theory," *International Conference on Communications*, 2018.

[7] S. Jin, W. Tan, M. Matthaiou, J. Wang, and K. Wong, "Statistical eigenmode transmission for the MU-MIMO downlink in rician fading," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6650–6663, 2015.

[8] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock, and T. Salzer, "Low complexity scheduling and beamforming for multiuser MIMO systems," *IEEE Workshop on Signal Processing Advances in Wireless Communications*, pp. 1 – 5, 2006.

[9] C. Zhang, Y. Huang, Y. Jing, S. Jin, and L. Yang, "Sum-rate analysis for massive MIMO downlink with joint statistical beamforming and user scheduling," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2181– 2194, 2017.

[10] D. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 502–513, Mar. 2000.

[11] M. Joham, K. Kusume, M. H. Gzara, W. Utschick, and J. A. Nossek, "Transmit wiener filter for the downlink of TDD DS-CDMA systems," *7th IEEE ISSSTA*, vol. 1, no. 1, pp. 9–13, Sept. 2002.

[12] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: Channel inversion and regularization," *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[13] V. K. Nguyen and J. S. Evans, "Multiuser transmit beamforming via regularized channel inversion: A large system analysis," *IEEE Global Telecommunications Conference*, pp. 1–4, Dec. 2008.

[14] R. Muharar and J. Evans, "Downlink beamforming with transmit-side channel correlation: A large system analysis," in *IEEE International Conference on Communications*, Jun. 2011, pp. 1–5.

[15] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[16] A. W. V. der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.

[17] R. Ertel, P. Cardieri, K. Sowerby, T. Rappaport, and J. Reed, "Overview of spatial channel models for anfenna array communication systems," *IEEE Personal Communications*, vol. 5, no. 1, pp. 10–22, Feb. 1998.