Full-Duplex vs. Half-Duplex: Delivery-Time Optimization in Cellular Downlink

Ali Kariminezhad, Soheil Gherekhloo and Aydin Sezgin

***Abstract*—The employment of the emerging trend of proactive caching leads to a trade-off between memory size and information rate. Due to the limited cache size only a limited number of demanded files might be available in the cache of the local server. Hence, the missing files need to be forwarded to the destinations. Now, the local server has the option to operate in either half-duplex or full-duplex mode. In half-duplex mode (HD), the demanded files are transmitted as soon as all files are provided to the local server. However, in a full-duplex mode (FD), the cached files can be communicated to the users, meanwhile receiving the missing files. Hence, by full-duplex operation, the local servers manage their cache (deplete/refill) in real-time adaptively which is beneficial from memory perspective. For both cases, i.e., HD and FD, the local server exploits decode-and-forward (DF) strategy. In this paper, we address these two strategies from the delivery-time perspective. As the worst link is the delivery-time bottleneck, fairness among users becomes of particular interest. Hence, we cast this problem as a min-max fair optimization problem. Moreover, we formulate the problem as a semi-definite relaxation (SDR) feasibility check problem after defining auxiliary variables to cope with non-convex constraints. The SDR comes along with bisection (over a single variable) and exhaustive search (over a few variables). Eventually, we compare the performance of half-duplex and full-duplex operations from the delivery-time perspective. Depending on the self-interference (SI) channel strength, the full-duplex operation can be beneficial from the delivery-time perspective up to almost 10% compared to the half-duplex counterpart.**

I. Introduction

Small-cell deployment has been reported to lead to improvements in spectral efficiency besides vast coverage in cellular networks [1]. Furthermore, utilizing multiple transmit chains provides the opportunity for serving multiple users with designated information, simultaneously [2]. Considering the main base station (BS) and local servers in the small-cells (LS), the communication burden can be moved from the BS to the LS by having all requested information available at the local servers (e.g. cached over night). However, this is not practical due to the absence of global information knowledge (e.g., in the caching phase, the demands are not known deterministically). Moreover, due to the cache size constraint, all files can not be cached at the local servers. In this paper, we consider the case that the information requested by some of the users are available at the local servers and some of the other are only available at the main base station. With the valid assumption that the links between the main base station and mobile stations (end users) are week or absent, the local servers have to act as relays in order to provide the information to the users.

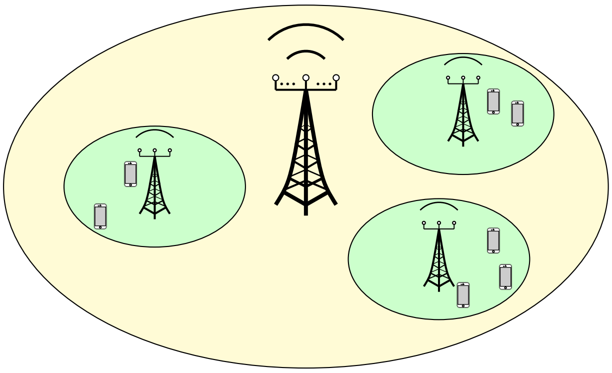

Fig. 1: Deploying several small cells in a macro cell. The transmitters are equipped with multiple transmit chains and the receivers have a single receive chain.

In this paper, we study the performance of two potential strategies for the LS (which operates as a relay with partial cognition). The LS can operate either in half-duplex mode or in full-duplex mode. The half-duplex LS first gathers all demanded information from the BS. Then, having all demands available, the LS starts transmission. The full-duplex LS, however, can transmit and receive simultaneously [3]. Notice that, the full-duplex operation requires extra signaling overhead for self-interference channel estimation [4] and further signal processing overhead for SI cancellation [5]–[7]. Full-duplex communication has been considered from different perspectives. For instance, the authors in [8] study the benefits of the full-duplex operation from the secrecy viewpoint, however [9] considers the energy efficiency of the full-duplex massive MIMO relaying. In this paper, we study the benefits of full-duplex operation from the delivery-time perspective. The authors in [10] study the fundamental information-theoretic limits on latency in small-cell caching. moreover, [11] addresses the trade-off between content caching at the local servers for reducing delivery-time and cloud computing for cooperation among local servers in expense of an increased delivery-time. These works mainly contribute on optimal caching-placement phase and the following communication phase at very high signal-to-noise power ratio (SNR).

Our contribution:

- Optimized transmission for achieving minimum delivery-time at arbitrary SNR with reliable information transmis-

A. Kariminezhad, Soheil Gherekhloo and A. Sezgin are with the Institute of Digital Communication Systems, Ruhr-Universität Bochum (RUB), Germany (emails: {ali.kariminezhad, soheyl.gherekhloo, aydin.sezgin}@rub.de).

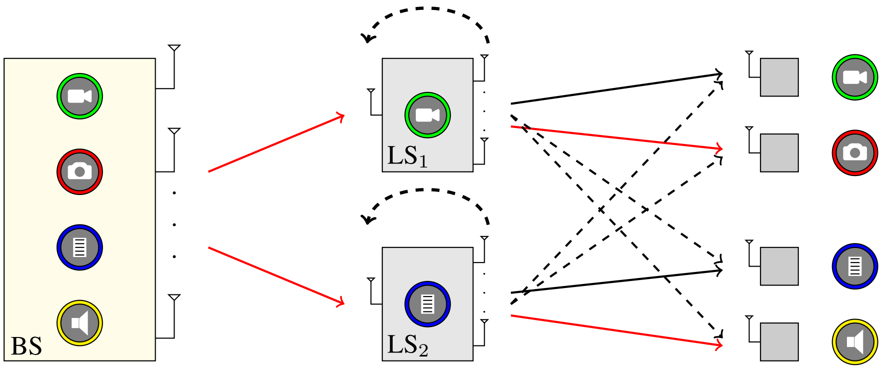

Fig. 2: Demanded files by the mobile stations in different small cells. Some of the demands are available at the LS, while the missing files need to be relayed by local servers using decode-and-forward (DF) relaying strategy.

sion.

- Application of full-duplex local servers in the post-caching scenario.
- Addressing two-fold benefits (memory and delivery-time), which is granted due to the full-duplex operation.

Here, the key element is to have fairness in serving the users. This is due to the fact that, minimum delivery-time is limited by the reliable information transmission rate over the worst link. Hence, optimized transmission in favour of the worst link, is delivery-time optimal. Therefore, we cast the problem as a min-max fair optimization problem, which can be formulated as a semidefinite program (SDP) with non-convex constraints. The non-convex terms are relaxed by a combinatorial bisection (a single auxiliary variable) and exhaustive search (one auxiliary variable per small cell) methods. By dropping rank-1 constraints, we obtain a convex feasibility check problem which is solved efficiently by interior-point methods (e.g., barrier method) [12]. Next, Gaussian randomization is applied to obtain feasible rank-1 sub-optimal solutions [13], [14].

Notation: We represent vectors in boldface lower-case letters while the matrices are expressed in boldface upper-case. $\text{tr}(\mathbf{A})$, $\mathbf{A}^H$ represent trace and hermitian of matrix $\mathbf{A}$, respectively. $\mathbf{B} = \text{diag}(\mathbf{A})$ is a diagonal matrix composed of the diagonal elements of matrix $\mathbf{A}$.

II. System Model

We consider a downlink cellular network as shown in Fig. 1. The macro-cell base station (BS) has N transmit antennas, while each small-cell local server (LS) is equipped with M transmit antennas and a single receive antenna (for simplification). We assume that, K small-cells are deployed in the macro-cell, hence K local servers cooperate with the macro-cell BS. Moreover, we consider single-antenna mobile stations (MS). Considering full-duplex, decode-and-forward LS, we model the channel input-output relationship as

$$y_k = \mathbf{h}_{\text{B}k}^H \mathbf{x}_\text{B} + \underbrace{\mathbf{h}_{kk}^H \mathbf{x}_k}_{\text{self-interference (SI)}} + z_k, \tag{1}$$

$$y_{kj} = \mathbf{h}_{kkj}^H \mathbf{x}_k + \sum_{\substack{l=1\\ l\neq k}}^{K} \mathbf{h}_{lkj}^H \mathbf{x}_l + z_{kj}, \tag{2}$$

where the received signal at the kth LS and at the jth MS in the kth cell are represented by $y_k \in \mathbb{C}$ and $y_{kj} \in \mathbb{C}$, respectively. The channel vector from the BS to the kth LS and from lth LS to the jth MS in the kth cell are depicted by $\mathbf{h}_{\text{B}k} \in \mathbb{C}^N$ and $\mathbf{h}_{lkj} \in \mathbb{C}^M$, respectively. Moreover, the self-interference (SI) channel is given by $\mathbf{h}_{kk} \in \mathbb{C}^M$. Here, we assume all channels are perfectly known at the BS and LS. The transmit signal from the BS and the kth LS, the receiver AWGN at the kth LS and jth MS in the kth cell are represented by $\mathbf{x}_\text{B} \in \mathbb{C}^N$, $\mathbf{x}_k \in \mathbb{C}^M$, $z_k \in \mathbb{C}$ and $z_{kj} \in \mathbb{C}$, respectively. The transmit signals from the BS and LS are composed of two parts, namely, the desired part (here we denote it by $\mathbf{s}_\text{B}$ and $\mathbf{s}_k$, respectively) and the transmitter noise (here denoted by $\mathbf{e}_\text{B}$ and $\mathbf{e}_k$, respectively) as

$$\mathbf{x}_\text{B} = \mathbf{s}_\text{B} + \mathbf{e}_\text{B}, \tag{3}$$

$$\mathbf{x}_k = \mathbf{s}_k + \mathbf{e}_k, \tag{4}$$

where the transmitter noise is mainly due to the amplification noise, limited dynamic range (DR) at the quantization phase. The information symbols are precoded at the multiple-antenna BS and LS before transmission. Here, we utilize linear precoding resulting in

$$\mathbf{s}_\text{B} = \sum_k \mathbf{v}_{\text{B}k} d_k, \tag{5}$$

$$\mathbf{s}_k = \sum_j \mathbf{u}_{kj} d_{kj}, \tag{6}$$

where d_k and d_{kj} are the unit-power independent information symbols. These information are designated for the kth LS for further relay processing and for the jth user in the kth small cell, respectively. The transmit directions are controlled by beamforming vectors, $\mathbf{v}_{\text{B}k}$ and $\mathbf{u}_{kj}$. It is important to notice that, the allocated transmit power for the designated messages for the kth LS and jth user in the kth small cell are represented by $\|\mathbf{v}_{\text{B}k}\|^2 = \text{tr}(\mathbf{C}_{\text{B}k})$ and $\|\mathbf{u}_{kj}\|^2 = \text{tr}(\mathbf{C}_{kj})$, respectively, where $\mathbf{C}_{\text{B}k} = \mathbf{v}_{\text{B}k}\mathbf{v}_{\text{B}k}^H$ and $\mathbf{C}_{kj} = \mathbf{u}_{kj}\mathbf{u}_{kj}^H$. Now, we define the covariance matrices of the desired part at the BS and kth

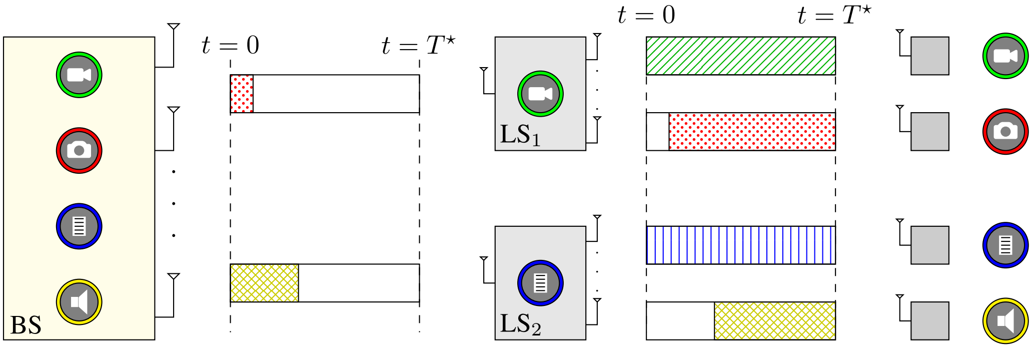

Fig. 3: The file transfer strategy from timing perspective while fairness is considered among the end users. Notice that, $T^\star$ is the optimal delivery-time

LS as

$$\mathbb{E}\{\mathbf{s}_\text{B}\mathbf{s}_\text{B}^H\} = \sum_k \mathbf{C}_{\text{B}k} = \mathbf{C}_\text{B}, \tag{7}$$

$$\mathbb{E}\{\mathbf{s}_k\mathbf{s}_k^H\} = \sum_j \mathbf{C}_{kj} = \mathbf{C}_k, \tag{8}$$

respectively. Furthermore, we utilize the following transmitter noise model, [15]

$$\mathbf{e}_\text{B} \perp\!\!\!\perp \mathbf{s}_\text{B}, \qquad \mathbf{e}_\text{B} \sim \mathcal{CN}(\mathbf{0}, \kappa \text{diag}(\mathbf{C}_\text{B})), \;\; \kappa \ll 1, \tag{9}$$

$$\mathbf{e}_k \perp\!\!\!\perp \mathbf{s}_k, \qquad \mathbf{e}_k \sim \mathcal{CN}(\mathbf{0}, \kappa \text{diag}(\mathbf{C}_k)), \;\; \kappa \ll 1. \tag{10}$$

which is based on the results in [16], [17]. Notice that, κ is defined as the transmit noise coefficient. Now, taking transmitter noise into account, the received signal is written as

$$y_k = \mathbf{h}_{\text{B}k}^H(\mathbf{s}_\text{B} + \mathbf{e}_\text{B}) + \mathbf{h}_{kk}^H(\mathbf{s}_k + \mathbf{e}_k) + z_k, \tag{11}$$

$$y_{kj} = \mathbf{h}_{kkj}^H(\mathbf{s}_k + \mathbf{e}_k) + \sum_{\substack{l=1\\ l\neq k}}^{K} \mathbf{h}_{lkj}^H(\mathbf{s}_l + \mathbf{e}_l) + z_{kj}. \tag{12}$$

In this paper, we assume that the information demands of all MSs are fully available at the BS, while few are cached at the LS serving the corresponding cell. For instance, as depicted in Fig. 2, the demands of the second users at each small cell are not locally available at the LS. However, the demand of the first users are locally available. To fulfil the demands of all users, the two following strategies are of interest,

- half-duplex LS (HD-LS): The LS either transmits to the mobile users or receives data from the BS. This way, the missing file is first fetched until having all demands available, then, LS starts transmission.
- full-duplex LS (FD-LS): The LS transmits cached requested files, meanwhile receiving the missing files from the BS. This way, self-interference is inevitable.

The resource allocation problem for the first strategy is trivial, while the second one requires further investigation. We present the solution for the second strategy, however, the comparison with the first strategy will appear in section IV.

Remark 1. *In point-to-point (P2P) communications, thermal noise at the receiver is mainly considered as the only barrier for the information rate given perfect channel state information (CSI). Notice that the transmitter noise is mainly ignored due to huge path loss from the source to the sink. However, for strong channels (here the SI channel), transmitter noise needs to be considered.*

Having the noise-free transmit signal at the LS and assuming perfectly known channel state information (CSI), the effect of self-interference (SI) can be partially cancelled [4]. Furthermore, we ignore the received transmitter noise at the end users due to relatively weak channels compared to the SI channel. Therefore, the received signals are formulated as

$$y_k = \mathbf{h}_{\text{B}k}^H\mathbf{s}_\text{B} + \mathbf{h}_{kk}^H\mathbf{e}_k + z_k, \tag{13}$$

$$y_{kj} = \mathbf{h}_{kkj}^H\mathbf{s}_k + \sum_{\substack{l=1\\ l\neq k}}^{K} \mathbf{h}_{lkj}^H\mathbf{s}_l + z_{kj}, \tag{14}$$

The received signals at the LS and end users consist of desired and interference parts as following

$$y_k = \underbrace{\mathbf{h}_{\text{B}k}^H\mathbf{s}_{\text{B}k}}_{\text{desired}} + \underbrace{\sum_{\substack{k'=1\\ k'\neq k}}^{K} \mathbf{h}_{\text{B}k}^H\mathbf{s}_{\text{B}k'}}_{\text{interference}} + \underbrace{\mathbf{h}_{kk}^H\mathbf{e}_k}_{\text{residual SI (RSI)}} + z_k, \tag{15}$$

$$y_{kj} = \underbrace{\mathbf{h}_{kkj}^H\mathbf{s}_{kj}}_{\text{desired}} + + \underbrace{\sum_{\substack{m=1\\ m\neq j}}^{|\mathcal{I}_k|+|\mathcal{I}_k^c|} \mathbf{h}_{kkj}^H\mathbf{s}_{km}}_{\text{intra-cell interference}} + \underbrace{\sum_{\substack{l=1\\ l\neq k}}^{K} \mathbf{h}_{lkj}^H\mathbf{s}_l}_{\text{inter-cell interference}} + z_{kj}, \tag{16}$$

where $\mathcal{I}_k$ denotes the set of mobile stations in the kth cell that demand the information files and the demanded files are in the cache of the respective LS. Notice that, $\mathcal{I}_k^c$ is the set of users whose information demands are fetched from the BS (i.e., the demands are not available at the kth LS). It is important to note that $\mathcal{I}_k^c \cap \mathcal{I}_k = \emptyset$.

Remark 2. *The studied system model and the relevant optimization problem capture the full-duplex operation only. As the half-duplex LS operates at orthogonal resources for transmission and reception, the system model and achievable information rates become trivial. Hence, HD-LS is not discussed in details due to page limit.*

$$R_{kj} = \log_2 \left(1 + \frac{\mathbf{h}_{kkj}^H \mathbf{C}_{kj} \mathbf{h}_{kkj}}{\sum_{\substack{m=1\\ m\neq j}}^{|\mathcal{I}_k|+|\mathcal{I}_k^c|} \mathbf{h}_{kkj}^H \mathbf{C}_{km} \mathbf{h}_{kkj} + \sum_{\substack{l=1\\ l\neq k}}^{K} \sum_{m=1}^{|\mathcal{I}_k|+|\mathcal{I}_k^c|} \mathbf{h}_{lkj}^H \mathbf{C}_{lm} \mathbf{h}_{lkj} + \sigma^2} \right), \quad \forall k, \forall j \in \mathcal{I}_k \cup \mathcal{I}_k^c, \tag{17}$$

$$R_{\mathrm{B}k} = \log_2 \left(1 + \frac{\mathbf{h}_{\mathrm{B}k}^H \mathbf{C}_{\mathrm{B}k} \mathbf{h}_{\mathrm{B}k}}{\sum_{\substack{k^{'}=1\\ k^{'}\neq k}}^{K} \mathbf{h}_{\mathrm{B}k}^H \mathbf{C}_{\mathrm{B}k'} \mathbf{h}_{\mathrm{B}k} + \kappa \mathbf{h}_{kk}^H \sum_{i=1}^{|\mathcal{I}_k|} \mathrm{diag}(\mathbf{C}_{ki}) \mathbf{h}_{kk} + \sigma^2} \right), \quad \forall k. \tag{18}$$

In this paper, we assume that all information contents (files) are of the same size L. Moreover, we assume that the demanded files are either not available at the LS or completely available. That means fractions of files are not considered. Hence, the duration of file transfer is restricted by the time that reliable information transfer is performed. Having the demanded file of the jth MS at the corresponding LS, the amount of time required for the file d_{kj} to be reliably transmitted to the jth MS in the kth cell is given by

$$t_{kj} = \frac{L}{BR_{kj}}, \quad j \in \mathcal{I}_k, \tag{19}$$

where B is the channel bandwidth in Hertz. Furthermore, R_{kj} is the achievable information rate in bits per channel use (here per second per hertz) which is formulated on the top of the page in (17), where σ^2 is the AWGN variance. Notice that, as shown in Fig. 3, the interference from the users in $\mathcal{I}_k^c$ should not appear in the received signal at the users in $\mathcal{I}_k$ for a small portion of time. Hence, the achievable rate in (17) serves as a lower-bound for the ultimate achievable rate due to an extra imposed interference for simplicity. Interestingly, we will still observe the superiority of the full-duplex operation with a lower-bounded rate compared to the half-duplex operation with the exact achievable rate (in section IV).

Now, the missing files need to be fetched from the BS. For that, the aggregate of the missing files of the users in kth cell (super file) is encoded and transmitted to the kth LS. Denoting the number of missing files in the kth cell as q_k, the BS-LS communication requires the total time of $t_{\mathrm{B}k} = \frac{Lq_k}{R_{\mathrm{B}k}}$ with the reliable transmission rate shown on the top of the page in (18). The amount of time required for passing the requested file to ith user for which the file is fetched from the BS is thus

$$t_{ki}^* = t_{\mathrm{B}k} + t_{ki}, \quad \forall i \in \mathcal{I}_k^c. \tag{20}$$

From (19) and (20), the total required time for passing the demanded files of size L to the users in the kth cell can be expressed as

$$T_k = \max\{t_{kj}, t_{ki}^*\}, \quad \forall j \in \mathcal{I}_k, \ i \in \mathcal{I}_k^c. \tag{21}$$

However, the delivery-time in the network can be written as

$$T = \max \ T_k. \tag{22}$$

It is worth mentioning that, the transmission delivery-time in (19) and (20) are functions of reliable information transmission rates which in turn are the functions of transmit covariance matrices. Henceforth, we aim at optimizing transmit covariance matrices at the BS and all LSs in order to minimize delivery-time in the whole files transfer phase.

Having the concept and definitions, we will formulate the delivery-time minimization problem in the next section.

III. Optimization Problem

Let P_{B} and P_k be the transmit power budget at the BS and kth LS, respectively. Furthermore, let the cone of Hermitian positive semidefinite matrices of size N and M be denoted by $\mathbb{H}_+^N$ and $\mathbb{H}_+^M$, respectively. Then, we formulate the network delivery-time minimization problem as

$$\begin{aligned} \min_{\mathbf{C}_{\mathrm{B}} \in \mathbb{H}_+^N, \ \mathbf{C}_k \in \mathbb{H}_+^M} \quad & \max_k \ T_k && (23) \\ \text{subject to} \quad & \mathrm{tr}(\mathbf{C}_{\mathrm{B}}) \leq P_{\mathrm{B}}, && (23a) \\ & \mathrm{tr}(\mathbf{C}_k) \leq P_k, \quad \forall k && (23b) \\ & \mathrm{rank}(\mathbf{C}_{km}) = 1, \quad \forall k, m \in \mathcal{I}_k \cup \mathcal{I}_k^c && (23c) \\ & \mathrm{rank}(\mathbf{C}_{\mathrm{B}k}) = 1, \quad \forall k && (23d) \end{aligned}$$

where the transmit covariance matrix at the transmitters should fulfil the rank-1 constraint expressed in (23c), (23d) in order to acquire a feasible beamforming solution. These constraints jointly with the objective function render the problem to a non-convex problem. However, optimal covariance solutions, i.e., $\mathbf{C}_{\mathrm{B}}^\star, \mathbf{C}_k^\star$, from problem (23) will eventually satisfy

$$t_{kj}(\mathbf{C}_{\mathrm{B}k}^\star, \mathbf{C}_{kj}^\star) \approx t_{ki}^*(\mathbf{C}_{\mathrm{B}k}^\star, \mathbf{C}_{kj}^\star), \quad \forall k \in \mathcal{K}, \tag{24}$$

$$T_k(\mathbf{C}_{\mathrm{B}k}^\star, \mathbf{C}_{kj}^\star) \approx T_{k'}(\mathbf{C}_{\mathrm{B}k}^\star, \mathbf{C}_{kj}^\star), \quad \forall k, k^{'} \in \mathcal{K}. \tag{25}$$

This is due to the optimal transmission for balancing the information delivery over unfair channels which frequently happens in wireless data transmission. Hence, fairness-optimal design yields equal file delivery-time for the demands in all small cells, Fig. 3. To this end, available resources (space and time) need to be allocated to the information demands (files) accordingly. Here, we enrich the optimization domain of problem (23) by defining scalar auxiliary variables as

$$\alpha_k = t_{\mathrm{B}k}, \ \forall k, \tag{26}$$

$$\beta_k = \max\{t_{kj}, t_{ki} + \alpha_k\}, \ \forall k, \tag{27}$$

$$\Gamma = \max_k \beta_k. \tag{28}$$

Now, optimization problem (23) is reformulated as

$$
\begin{aligned}
\min_{\Gamma,\beta,\alpha,\ \mathbf{C}_{\mathrm{B}},\ \mathbf{C}_k} \quad & \Gamma && (29)\\
\text{subject to} \quad & \Gamma \geq \beta_k, \quad \forall k, && (29a)\\
& \beta_k \geq t_{kj}, \quad \forall k, j \in \mathcal{I}_k, && (29b)\\
& \beta_k \geq t_{ki} + \alpha_k, \quad \forall k, i \in \mathcal{I}_k^c, && (29c)\\
& \alpha_k \geq t_{\mathrm{B}k}, \quad \forall k, && (29d)\\
& \mathrm{tr}(\mathbf{C}_{\mathrm{B}}) \leq P_{\mathrm{B}}, && (29e)\\
& \mathrm{tr}(\mathbf{C}_k) \leq P_k, \quad \forall k, && (29f)\\
& \mathrm{rank}(\mathbf{C}_{km}) = 1, \quad \forall k, m \in \mathcal{I}_k \cup \mathcal{I}_k^c && (29g)\\
& \mathrm{rank}(\mathbf{C}_{\mathrm{B}k}) = 1, \quad \forall k. && (29h)
\end{aligned}
$$

By combining (29a)-(29c) and bisecting over the auxiliary scalar variable Γ, we obtain a non-convex feasibility check problem, which is

$$
\begin{aligned}
\text{find} \quad & \alpha_k \in \mathbb{R},\ \mathbf{C}_{\mathrm{B}k} \in \mathbb{H}_+^N,\ \mathbf{C}_{km} \in \mathbb{H}_+^M\ \forall k && (30)\\
\text{subject to} \quad & R_{kj} \geq {}^{L}\!/\Gamma, \quad \forall k, j \in \mathcal{I}_k, && (30a)\\
& R_{ki} \geq {}^{L}\!/(\Gamma - \alpha_k) \quad \forall k, i \in \mathcal{I}_k^c && (30b)\\
& R_{\mathrm{B}k} \geq {}^{Lq_k}\!/\alpha_k, \quad \forall k, && (30c)\\
& \mathrm{tr}(\mathbf{C}_{\mathrm{B}}) \leq P_{\mathrm{B}}, && (30d)\\
& \mathrm{tr}(\mathbf{C}_k) \leq P_k, \quad \forall k && (30e)\\
& \mathrm{rank}(\mathbf{C}_{km}) = 1, \quad \forall k, m \in \mathcal{I}_k \cup \mathcal{I}_k^c && (30f)\\
& \mathrm{rank}(\mathbf{C}_{\mathrm{B}k}) = 1, \quad \forall k. && (30g)
\end{aligned}
$$

The constraint set in (30) is a non-convex set. This is due to constraints (30b) and (30c). Besides, rank constraints in (30f) and (30g) are non-convex constraints as well. This problem is difficult (can not be solved in polynomial time). Here, we define $\mathbf{H}_{kkj} = \mathbf{h}_{kkj}\mathbf{h}_{kkj}^H$, $\mathbf{H}_{\mathrm{B}k} = \mathbf{h}_{\mathrm{B}k}\mathbf{h}_{\mathrm{B}k}^H$, and $\mathbf{H}_{kk} = \mathbf{h}_{kk}\mathbf{h}_{kk}^H$, and take trace operation from the numerator and denominator of the SINR terms in the rate expressions of (17) and (18). Then, interestingly for given $\alpha_k,\ \forall k$, the semidefinite relaxation (SDR) of problem (30) is a convex program. Hence, we solve the feasibility check problem in (32) for given $\alpha_k,\ \forall k$ and Γ. Therefore, we exhaustively search over $\alpha_k,\ \forall k$, bisect over Γ and solve the following feasibility check problem for each step in the exhaustive search,

$$
\begin{aligned}
\text{find} \quad & \mathbf{C}_{\mathrm{B}} \in \mathbb{H}_+^N,\ \mathbf{C}_k \in \mathbb{H}_+^M\ \forall k && (32)\\
\text{subject to} \quad & R_{kj} \geq {}^{L}\!/\Gamma, \quad \forall k, j \in \mathcal{I}_k, && (32a)\\
& R_{ki} \geq {}^{L}\!/(\Gamma - \alpha_k) \quad \forall k, i \in \mathcal{I}_k^c && (32b)\\
& R_{\mathrm{B}k} \geq {}^{Lq_k}\!/\alpha_k, \quad \forall k, && (32c)\\
& \mathrm{tr}(\mathbf{C}_{\mathrm{B}}) \leq P_{\mathrm{B}}, && (32d)\\
& \mathrm{tr}(\mathbf{C}_k) \leq P_k, \quad \forall k && (32e)
\end{aligned}
$$

It is important to note that, the solution of the problem does not necessarily fulfil the rank-1 constraint. Hence, a rank reduction procedure (e.g., Gaussian randomization [13], [14]) might be required if the solutions for covariance matrices are well-conditioned. However, in case that the solutions have one significantly dominant eigen direction (i.e. $\frac{\text{maximum eigen-value}}{\text{minimum eigen-value}}$ is sufficiently large), we get a single rank solution based on the dominant direction by eigen-value decomposition.

Needless to say that, exhaustive search over $\alpha_k \in \mathbb{R}\ \forall k$, is not an efficient approach, since the feasibility check problem needs to be solved for each $\alpha_k \in \mathbb{R}\ \forall k$, with a pre-defined resolution.

Algorithm 1 Determine α_{min}

```
Zero force SI at LS precoder: $\mathbf{v}_{kj} = \mathrm{Null}\{\mathbf{H}_{kk}\}\ \forall k, j \in \mathcal{I}_k$,
Notice, $R_{\mathrm{B}k} = \log_2(1 + \frac{\mathbf{h}_{\mathrm{B}k}^H \mathbf{C}_{\mathrm{B}k} \mathbf{h}_{\mathrm{B}k}}{\sum_{k' \neq k}^{K} \mathbf{h}_{\mathrm{B}k}^H \mathbf{C}_{\mathrm{B}k'} \mathbf{h}_{\mathrm{B}k} + \sigma^2}), \quad \forall k$,
Define the auxiliary variable $\Theta = \max_k(t_{\mathrm{B}k})$,
consider the following feasibility check problem
    $\text{find} \quad \mathbf{C}_{\mathrm{B}k} \in \mathbb{H}_+^N,\ \forall k$    (31)
    $\text{subject to} \quad R_{\mathrm{B}k} \geq {}^{Lq_k}\!/\Theta, \quad \forall k$,    (31a)
    $\sum_{k=1}^{K} \mathrm{tr}(\mathbf{C}_{\mathrm{B}_k}) \leq P_{\mathrm{B}}$,    (31b)
Determine $\Theta = \Theta_{min} \rightarrow 0$ for which problem (31) is not feasible,
Determine $\Theta = \Theta_{max}$ (very large) for which problem (31) is feasible,
Solve (31) for $\Theta^{(l)} = \frac{\Theta_{min} + \Theta_{max}}{2}$, $l = 1$ (iteration index),
if solution exists, then
    solve (31) for $\Theta^{(l+1)} = \frac{\Theta_{min} + \Theta^{(l)}}{2}$,
    update $\Theta_{max} = \Theta^{(l+1)}$,
else
    solve (31) for $\Theta^{(l+1)} = \frac{\Theta_{max} + \Theta^{(l)}}{2}$,
    update $\Theta_{min} = \Theta^{(l+1)}$,
end if
Iteration ends when specified resolution $\epsilon = \Theta^{l+1} - \Theta^{l}$ is achieved,
The solution $\mathbf{C}_{\mathrm{B}k}$ are rank-1, hence no need for rank reduction procedure,
Determine $\alpha_{min} = \Theta^{(\mathrm{end})}$.
```

Algorithm 2 Network delivery-time Optimization

```
for $\alpha_k = \alpha_{min}$ : resolution : $\alpha_{max},\ \forall k$ do
    $l = l + 1$ (iteration index),
    Bisection over $\Gamma$, similar to Algorithm 1,
    Solve problem (32),
    Perform rank reduction procedure,
    Result: $\Gamma^\star(l)$, $\mathbf{C}_{\mathrm{B}k}(l)\ \forall k$, $\mathbf{C}_{kj}(l)\ \forall k, j$ ,
end for
delivery-time= $\min_l\ \Gamma^\star$,
$l^\star = \arg\min_l\ \Gamma^\star$,
Optimal transmit covariance matrices: $\mathbf{C}_{\mathrm{B}k}^\star(l^\star)$, $\mathbf{C}_{kj}^\star(l^\star)$.
```

A. Exhaustive search domain

Here, we limit the exhaustive search domain in order to decrease the number of feasibility checks (solving problem (32)). Hence, we limit $\alpha_k \in \mathbb{R}$ from below by α_{min} and from above by α_{max}.

- α_{min} specifies the minimum delivery-time of the missing files from the BS to the kth LS under transmit power constraints. This delivery-time is minimized if the beamforming strategy at the kth LS steers the transmit signal to the jth mobile station ($j \in \mathcal{I}_k$) onto the null-space of the SI channel. This way, the problem simplifies to a broadcast channel and the minimum delivery-time can be determined by solving a SDR which yields a rank-1 solution intrinsically [18]. The procedure of determining α^{min} is explained in Algorithm 1.

- α_{max} determines the maximum file delivery-time from the MSB to LS. This bound is determined as

$$\alpha_{max} = \alpha_{min} + \zeta \tag{33}$$

where ζ is the minimum amount of time required for files transfer in the small cells, while providing the missing files in a genie-aided fashion to the corresponding local server (LS). This problem also simplifies to the beamforming problem of an interference broadcast channel (IBC), which can be formulated as a SDR and solved by interior point methods and further Gaussian randomization to acquire a rank-1 solution.

Having determined α_{min} and α_{max}, Algorithm 2 gives the delivery-time optimization problem, elaborately.

IV. Numerical Results

In this section we present the numerical results for the full-duplex relaying from the delivery-time perspective. Furthermore, as a benchmark we compare the performance gain of full-duplex operation with the half-duplex counterpart. Here, we consider $N = 10$ (number of antennas at the macro-cell base station) and $M = 4$ (number of antennas at the small cell local servers). Moreover, we consider two small-cells with two active mobile stations. The demanded file of one of the mobile stations is available at LS, while the demand of the other needs to be fetched from the BS. The files length is set to 5 information bits, i.e., $L = 5$. We consider the transmitter noise coefficient to be $\kappa = 0.1$. The numerical results based on Algorithm 1 and Algorithm 2 are depicted in Tab. I. According to this table, full-duplex relaying is capable of decreasing the delivery-time up to 10% compared to half-duplex operation. Notice that this improvement depends on the given strength of the self-interference channel.

V. Conclusion

In this paper, we studied the potential strategies at the local server with missing files in cellular downlink communication. The min-max fair optimization problem is utilized for minimizing the delivery-time of the worst link in the downlink channel. To achieve this goal, transmit beamforming directions jointly with the power allocation at each direction are subject to optimization. However, power constraint at the transmitters (the macro-cell base station and the local servers) need to be fulfilled. This problem with the specified constraints turned out to be a non-convex optimization problem, hence difficult to solve. By defining auxiliary variables and utilizing the relaxation method (SDR), we proposed an algorithm to capture the delivery-time of the downlink for full-duplex relaying. Finally, by comparing half-duplex and full-duplex operations, the superiority of full-duplex operation is emphasized from delivery-time perspective. However, full-duplex operation provides the opportunity for real-time adaptive cache management, which is advantages from the memory-size perspective.

SI Channel Realizations	Half-Duplex	Full-Duplex	Imp. %
Strong SI channel	9.2930 sec	8.9459 sec	3.7%
Moderate SI channel	9.2930 sec	8.5073 sec	8.4%
Weak SI channel	9.2930 sec	8.3763 sec	9.9%
No RSI	9.2930 sec	8.2914 sec	10.7%

TABLE I: Performance comparison of half-duplex and full-duplex operation from the delivery-time perspective for different SI channel realizations. The amount of 5 bits per channel bandwidth unit (Hertz) can be reliably transmitted with the above-mentioned delivery-time.

References

[1] A. Laya, K. Wang, A. A. Widaa, J. A.-Zarate, J. Markendahl, and L. Alonso, "Device-To-Device Communications And Small Cells: Eabling Spectrum Reuse For Dense Networks," *IEEE Wireless Communication Magazine*, August 2014.

[2] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2239–2251, June 2006.

[3] D. W. Bliss, P. A. Parker, and A. R. Margetts, "Simultaneous Transmission and Reception for Improved Wireless Network Performance," in *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, Aug 2007, pp. 478–482.

[4] A. Masmoudi and T. Le-Ngoc, "Channel Estimation and Self-Interference Cancellation in Full-Duplex Communication Systems," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2016.

[5] E. Everett, A. Sahai, and A. Sabharwal, "Passive Self-Interference Suppression for Full-Duplex Infrastructure Nodes," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 680–694, February 2014.

[6] A. Elsayed and A. Eltawil, "All-Digital Self-Interference Cancellation Technique for Full-Duplex Systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3519–3532, July 2015.

[7] H. Vogt, K. Ramm, and A. Sezgin, "Practical Secret-Key Generation by Full-Duplex Nodes with Residual Self-Interference," in *WSA 2016; 20th International ITG Workshop on Smart Antennas*, March 2016, pp. 1–5.

[8] H. Vogt and A. Sezgin, "Full-duplex vs. half-duplex secret-key generation," in *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, Nov 2015, pp. 1–6.

[9] H. Q. Ngo, H. A. Suraweera, M. Matthaiou, and E. G. Larsson, "Multipair Full-Duplex Relaying With Massive Arrays and Linear Processing," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 9, pp. 1721–1737, Sept 2014.

[10] S. M. Azimi, O. Simeone, and R. Tandon, "Fundamental Limits on Latency in Small-Cell Caching Systems: An Information-Theoretic Analysis," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[11] A. Sengupta, R. Tandon, and O. Simeone, "Cloud RAN and edge caching: Fundamental performance trade-offs," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–5.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[13] N. Sidiropoulos, T. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2239–2251, June 2006.

[14] Z. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, "Approximation Bounds for Quadratic Optimization with Homogeneous Quadratic Constraints," *SIAM Journal on Optimization*, vol. 18, no. 1, pp. 1–28, 2007.

[15] B. Day, A. Margetts, D. Bliss, and P. Schniter, "Full-Duplex MIMO Relaying: Achievable Rates Under Limited Dynamic Range," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 8, pp. 1541–1553, September 2012.

[16] H. Suzuki, T. V. A. Tran, I. Collings, G. Daniels, and M. Hedley, "Transmitter Noise Effect on the Performance of a MIMO-OFDM Hardware Implementation Achieving Improved Coverage," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 867–876, August 2008.

[17] G. Santella and F. Mazzenga, "A hybrid analytical-simulation procedure for performance evaluation in M-QAM-OFDM schemes in presence of nonlinear distortions," *IEEE Trans. Veh. Tech.*, vol. 47, no. 1, pp. 142–151, Feb 1998.

[18] X. Shang, B. Chen, and H. V. Poor, "Multiuser MISO Interference Channels With Single-User Detection: Optimality of Beamforming and the Achievable Rate Region," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4255–4273, July 2011.