Linear Precoding Design for Cache-aided Full-duplex Networks

Thang X. Vu*, Trinh Anh Vu†, Lei Lei*, Symeon Chatzinotas*, and Björn Ottersten*
* Interdisciplinary Centre for Security, Reliability and Trust (SnT) – University of Luxembourg, 29 Av. J.F. Kennedy, L-1855 Luxembourg. Email: {thang.vu, lei.lei, symeon.chatzinotas, bjorn.ottersten}@uni.lu.
† Dept. of Electronics and Telecommunications, VNU University of Engineering and Technology, Hanoi, Vietnam. Email: vuta@vnu.edu.vn.

***Abstract*—Edge caching has received much attention as a promising technique to overcome the stringent latency and data hungry challenges in the future generation wireless networks. Meanwhile, full-duplex (FD) transmission can potentially double the spectral efficiency by allowing a node to receive and transmit simultaneously. In this paper, we study a cache-aided FD system via delivery time analysis and optimization. In the considered system, an edge node (EN) operates in FD mode and serves users via wireless channels. Two optimization problems are formulated to minimize the largest delivery time based on the two popular linear beamforming zero-forcing and minimum mean square error designs. Since the formulated problems are non-convex due to the self-interference at the EN, we propose two iterative optimization algorithms based on the inner approximation method. The convergence of the proposed iterative algorithms is analytically guaranteed. Finally, the impacts of caching and the advantages of the FD system over the half-duplex (HD) counterpart are demonstrated via numerical results.**

***Index terms*—** Edge caching, delivery time, full-duplex, optimization.

I. Introduction

Among potential enabling technologies to tackle with stringent latency and data hungry challenges in future wireless networks, edge caching has received much attention. By prefetching content closer to end users at the edge node's local storage, edge caching can significantly reduce transmission latency and backhaul's traffic since the edge node can directly serve the users' demands without requesting for data transfer from the core network. [1]. Joint design for content caching and physical layer has attracted much attention recently. The main idea is to take into account the cached content at the edge nodes when designing the signal transmission to reduce costs on both access and backhaul links. Since some (parts of) the requested files are available in the edge node's cache, proper design is required for content selection combined with broad/multi-cast transmission design to improve the system performance, including energy efficiency [2], [3], throughput-outage tradeoff [4], and delivery time [5]. The performance of cache-aided wireless networks can be further improved by joint optimization of caching along with routing and resource allocation [6].

Meanwhile, full-duplex (FD) has shown great potential as the transmission technique for the next generation wireless networks. Thanks to recent developments in self-interference cancellation, FD can potentially double the spectral efficiency by allowing a node to transmit and receive signals simultaneously [7]. The employment of FD systems with caching capability comes as a step forward to further improve the system performance. It is shown via stochastic geometry analysis that a cache-aided FD system can positively provide cache hit enhancements compared with the half-duplex (HD) mode in heterogeneous networks (HetNets) [8] and device-to-device (D2D) systems [9], [10]. The worse case normalized delivery time (NDT) in HetNets is studied in [11] with FD relaying nodes. However, the results in [11] are based on an optimistic assumption that self-interference can be fully mitigated. In practice, there always remains residual interference after the self-interference cancellation [12], [13].

In this paper, we investigate the delivery time performance of a cache-aided FD system by taking into consideration realistic self-interference cancellation modelling. Our goal is to minimize the average delivery time via joint precoding vectors design for both backhaul and access links, which is fundamentally different from [8–10]. Two delivery time minimization problems are formulated based on the two popular linear beamforming zero-forcing (ZF) and minimum mean square error (MMSE) designs. To cope with the non-convexity of the formulated problems, two iterative optimization algorithms are proposed based on the inner approximation method. The convergence of the proposed iterative algorithms are analytically guaranteed. Finally, numerical results are presented to demonstrate the advantages of the proposed algorithms over the half-duplex system in certain scenarios.

Notation: $(.)^H, (.)^T$ and $(.)^{-1}$ denote the conjugate operator, transpose operator, and the inverse matrix, respectively.

The rest of this paper is organised as follows. Section II presents the system model and the caching strategies. Section II-B1 presents the delivery time optimization for the ZF design. Section IV optimizes the delivery time based on the MMSE design. Numerical results are shown in Section V. Finally, Section VI provides conclusions and discussions.

II. System Model and Performance Metric

We consider a cache-aided FD system, in which an edge node (EN) operates in FD mode and connects to the core network via a wireless backhaul access point (WAP), e.g., high power high tower or macro base station, as depicted in Fig. 1. The users can only access data from the EN via wireless access channels, i.e., there is no direct link between the users and the WAP. The WAP is assumed to have access to a library of F contents, denoted by $\mathcal{F} = \{f_1, \ldots, f_N\}$. Without loss of generality, all content is assumed to have equal size of Q

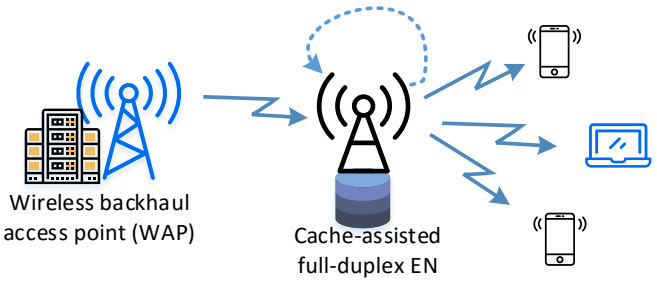

Fig. 1: Cache-aided full-duplex network. The edge node operates in full-duplex mode, while the users and backhaul wireless access point operate in half-duplex mode. Self interference occurs at the edge node.

bits. To leverage the backhaul during peak-hours, the EN is equipped with a storage memory of MQ bits, where $M < F$.

A. Content popularity and caching model

We consider the most popular content popularity model, i.e., the Zipf distribution. The probability for file f_n being requested is equal to

$$\nu_n = \Gamma^{-1} n^{-\xi}, \tag{1}$$

where $\Gamma = \sum_{m=1}^{F} m^{-\xi}$ and ξ is the Zipf skewness factor.

We consider generic caching policy $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_F\}$, where $\mu_n \in [0, 1]$ denotes parts of file f_n cached at the EN. In order to meet the memory constraint, it must hold that $\sum_{n=1}^{F} \mu_n \leq M$. The motivation behind the generic caching policy is that it allows to study different caching strategies. For the most popular (Zipf-based) caching, we have $\boldsymbol{\mu}_{\text{Zip}} = [\underbrace{1, \ldots, 1}_{\times M}, 0, \ldots, 0]$.

B. Signal transmission model

Let L, N denote the number of antennas at the WAP and the EN, respectively; $\boldsymbol{G} \in \mathbb{C}^{N \times L}$ denote the backhaul channel fading coefficients, including the path loss, whose elements are identically independently distributed (i.i.d.) complex Gaussian variables with zero mean and variance σ_{bh}^2; and $\boldsymbol{h}_k \in \mathbb{C}^{1 \times N}$ denote the access channel fading coefficients between the EN and user k, including the path loss, whose elements are i.i.d. complex Gaussian random variables with zero mean and variance σ_{ac}^2. Furthermore, denote $\boldsymbol{G}_0 \in \mathbb{C}^{N \times N}$ as the self-interference coefficients at the EN. Full channel state information is assumed to be known at the transmitter.

1) Transmission on backhaul link: When a user requests a content, it sends the content index to the EN. If (portions of) the requested content is available in the cache, it serves the user directly via the access channel. Otherwise, the EN will demand the non-cached parts from the WAP via the wireless backhaul before serving the user.

Denote $\boldsymbol{d} = [d_1, \ldots, d_K]$ as the request file indexes from the users. We consider the worst case when the users request K different files[1]. Under the caching policy $\boldsymbol{\mu}$, μ_{d_k} portions of the requested file f_{d_k} are already available in the EN's cache. Thus, the WAP needs only send the $1 - \mu_{d_k}$ non-cached parts of file f_{d_k} on the backhaul. Let s_k denote the modulated signal of the non-cached parts of file f_{d_k}, and denote $\boldsymbol{s} = [s_1, \ldots, s_K]$ as the aggregated signal sent through the backhaul. The received signal at the EN is given as

$$\boldsymbol{y}_E = \boldsymbol{G}\boldsymbol{s} + \boldsymbol{G}_0\boldsymbol{x} + \boldsymbol{n}_E, \tag{2}$$

where $\boldsymbol{x}$ is the EN's transmit signal which will be described in Sec. II-B2, the second term in (2) represents the self-interference at the EN due to the FD transmission, and $\boldsymbol{n}_E$ is the noise vector whose elements are complex Gaussian variables with zero mean and variance σ^2.

In order to decode $\boldsymbol{y}_E$, the EN first eliminates the self interference, since $\boldsymbol{x}$ is already known. After interference cancellation, the residual interference power is ηP_{EN}, where $P_{EN} = \| \boldsymbol{x} \|^2$ is the transmit power at the EN and η represents the interference cancellation efficiency. The common value of η is between -40dB and -80dB depending on the hardware and interference cancellation techniques [12], [13]. The achievable information rate on the backhaul link, by treating the self-interference as noise, is given as

$$\begin{aligned} C_{BH} =& W \log_2 \left(\det \left(\boldsymbol{I} + \frac{\boldsymbol{G}^H \Sigma_s \boldsymbol{G}}{\eta P_{EN} + \sigma^2} \right) \right) \\ =& \sum_{l=1}^{\bar{L}} W \log_2 \left(1 + \frac{\lambda_l q_l}{\eta P_{EN} + \sigma^2} \right), \end{aligned} \tag{3}$$

where W is the channel bandwidth, λ_l and $\bar{L} \leq \min(L, N)$ are the l-th eigenvalue and the rank of matrix $\boldsymbol{G}^H \boldsymbol{G}$, respectively; q_l is the power allocated for the l-th sub backhaul channel; and $\Sigma_s = \texttt{diag}(q_1, \ldots, q_{\bar{L}})$.

We employ the frequency division multiplexing access (FDMA) to allocate the backhaul capacity for the user requests. The backhaul capacity for user k is

$$C_k = \rho_k C_{BH} = \rho_k W \sum_{l=1}^{\bar{L}} \log_2 \left(1 + \frac{\lambda_l q_l}{\eta P_{EN} + \sigma^2} \right), \tag{4}$$

where $\rho_k = \frac{\bar{\mu}_k}{\sum_{k=1}^{K} \bar{\mu}_k}$, with $\bar{\mu}_k \triangleq 1 - \mu_{d_k}$.

2) Transmission on the access links: Let x_k denote the modulated signal of f_{d_k} targeting user k and $\boldsymbol{x} = \sum_{k=1}^{K} \boldsymbol{w}_k x_k$ denote the transmit signal at the EN, where $\boldsymbol{w}_k \in \mathbb{C}^{N \times 1}$ is the precoding vector for user k. The received signal at user k from the EN is given as

$$y_{U,k} = \boldsymbol{h}_k \boldsymbol{w}_k x_k + \sum_{i \neq k} \boldsymbol{h}_k \boldsymbol{w}_i x_i + n_{U,k}, \tag{5}$$

where $n_{U,k}$ is the Gaussian noise with zero mean and variance σ^2. The first term in (5) is the desired signal for user k, and the second term represents the interference from other users' information. The achievable information rate for user k, by treating interference as noise, is given as

$$R_k = W \log_2 \left(1 + \frac{|\boldsymbol{h}_k \boldsymbol{w}_k|^2}{\sum_{i \neq k} |\boldsymbol{h}_k \boldsymbol{w}_i|^2 + \sigma^2} \right). \tag{6}$$

The total transmit power at the EN is $P_{EN} = \| \boldsymbol{x} \|^2 = \sum_{k=1}^{K} \| \boldsymbol{w}_k \|^2$.

[1]This happens with high probability when K is small compared with F, which is usually true in practice.

In this paper, we consider two popular linear precodings ZF and MMSE due to their low computational complexity. The unified expression for the linear precoder is as:

$$\boldsymbol{w}_k = \begin{cases} \sqrt{p_k}\tilde{\boldsymbol{h}}_k, & \text{if ZF} \\ \sqrt{p_k}\check{\boldsymbol{h}}_k, & \text{if MMSE} \end{cases}, \tag{7}$$

where p_k is the power factor allocated for user k; $\tilde{\boldsymbol{h}}_k$ is the ZF beamforming vector, which is the k-th column of the ZF precoding matrix $\boldsymbol{H}^H(\boldsymbol{H}\boldsymbol{H}^H)^{-1}$; and $\check{\boldsymbol{h}}_k$ is the MMSE beamforming vector, which is the k-th column of the MMSE precoding matrix $\boldsymbol{H}^H(\sigma^2\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^H)^{-1}$, with $\boldsymbol{H} = [\boldsymbol{h}_1^T, \ldots, \boldsymbol{h}_K^T]^T$. In the following, we propose an optimization algorithm to minimize the delivery time under these two precoding methods.

III. Delivery time minimization under ZF design

In this section, we propose an optimal power allocation to minimize the delivery time based on the ZF beamforming. Note that under the ZF design, we have $\boldsymbol{h}_k\tilde{\boldsymbol{h}}_i = \delta_{ki}$, i.e., the inter-user interference is fully cancelled out. Therefore, the achievable rate on the access link for user k is

$$R_{ZF,k} = W\log_2\left(1 + \frac{p_k}{\sigma^2}\right), \tag{8}$$

and the total transmit power at the EN is $P_{EN} = \sum_{k=1}^{K}\alpha_k p_k$, where $\alpha_k \triangleq \| \tilde{\boldsymbol{h}}_k \|^2$.

The EN employs FastForward FD transmission [14], in which the delay of the forward signal is within the cyclic prefix (CP) duration. Therefore, the delivery time for the k-th user's request is $t_k = \frac{Q}{R_{ZF,k}}$ subjected to a condition that the EN's buffer is not empty. Because $\mu_{d_k}Q$ bits of the requested file is already in the EN's cache, this condition reads $C_k\tau + \mu_{d_k}Q \geq R_{ZF,k}\tau, \forall\tau \in [0, t_k]$. Consider all possible values of $\tau \in [0, t_k]$, this constraint becomes $C_k \geq (1-\mu_{d_k})R_{ZF,k} = \bar{\mu}_k R_{ZF,k}$, where $\bar{\mu}_k \triangleq 1 - \mu_{d_k}$.

We would like to minimize the largest delivery time among the users. The optimization problem is formulated as follows:

$$\underset{\{p_k, q_l\}}{\texttt{minimize}} \quad \max\left(\frac{Q}{R_{ZF,1}}, \ldots, \frac{Q}{R_{ZF,K}}\right), \tag{9}$$

$$\texttt{s.t.} \quad C_k \geq \bar{\mu}_k R_{ZF,k}, \forall k \tag{9a}$$

$$\sum_{k=1}^{K}\alpha_k p_k \leq P_\Sigma^{EN};\ \sum_{l=1}^{\bar{L}} q_l \leq P_\Sigma^{BS}, \tag{9b}$$

where P_Σ^{BS} and P_Σ^{EN} are the maximum transmit power at the WAP and the EN, respectively, and $\{p_k, q_l\}$ is the short-hand notation for the sets $\{p_k\}_{k=1}^K, \{q_l\}_{l=1}^{\bar{L}}$.

By introducing an arbitrary positive variable t and using (8) and (4), the problem (9) is equivalent to the following:

$$\underset{t, \{p_k, q_l\}}{\texttt{minimize}} \quad t \tag{10}$$

$$\texttt{s.t.} \quad \log\left(1 + \frac{p_k}{\sigma^2}\right) \geq \frac{Q\log(2)}{tW}, \forall k \tag{10a}$$

$$\rho_k\sum_{l=1}^{\bar{L}}\log\left(1 + \frac{\lambda_l q_l}{\eta\sum_{i=1}^{K}\alpha_i p_i + \sigma^2}\right) \geq \bar{\mu}_k\log\left(1 + \frac{p_k}{\sigma^2}\right), \forall k \tag{10b}$$

$$\sum_{k=1}^{K}\alpha_k p_k \leq P_\Sigma^{EN};\ \sum_{l=1}^{\bar{L}} q_l \leq P_\Sigma^{BS}. \tag{10c}$$

It is evident that problem 10 is non-convex due to constraint (10b). To overcome this difficulty, we will express this constraint into a convex expression. Denote $A \triangleq [\eta\alpha_1, \ldots, \eta\alpha_K]$, and $\boldsymbol{p} = [p_1, \ldots, p_K]^T$ as the compact form of the EN's transmit power vector. Then we can reformulate problem (10) as follows:

$$\underset{t, \{p_k, q_l\}}{\texttt{minimize}} \quad t \tag{11}$$

$$\texttt{s.t.} \quad \text{(10c)};\ \log\left(1 + \frac{p_k}{\sigma^2}\right) \geq \frac{Q\log(2)}{tW}, \forall k \tag{11a}$$

$$\rho_k\sum_{l=1}^{\bar{L}}\log(A\boldsymbol{p} + \lambda_l q_l + \sigma^2) \geq \bar{\mu}_k\log\left(1 + \frac{p_k}{\sigma^2}\right) + \rho_k\bar{L}\log(A\boldsymbol{p} + \sigma^2), \forall k \tag{11b}$$

where the constraint (11b) is obtained since $A\boldsymbol{p}+\sigma^2$ is strictly positive.

It is observed that problem (11) is non-convex since the second constraint is non-affine. By introducing arbitrary variables $\{x_k\}_{k=1}^K$ and y, we can reformulate problem (11) as

$$\underset{t, \{p_k, q_l, x_k\}, y}{\texttt{minimize}} \quad t \tag{12}$$

$$\texttt{s.t.} \quad \text{(10c)};\ \log\left(1 + \frac{p_k}{\sigma^2}\right) \geq \frac{Q\log(2)}{tW}, \forall k \tag{12a}$$

$$\rho_k\sum_{l=1}^{\bar{L}}\log(A\boldsymbol{p} + \lambda_l q_l + \sigma^2) \geq \bar{\mu}_k x_k + y, \forall k \tag{12b}$$

$$A\boldsymbol{p} \leq e^y;\ 1 + p_k/\sigma^2 \leq e^{x_k}, \forall k. \tag{12c}$$

Although constraints (12a) and (12b) are now convex, solving problem (12) is still challenging since constraint (12c) is unbounded. Fortunately, because the function e^x is convex, we can employ the inner approximation method, which uses the first-order approximation of the exponential function in the right hand side of constraint (12c). The approximated problem is stated as follows:

$$\boldsymbol{Q}_1(\boldsymbol{x}_0, y_0): \underset{t, \{p_k, q_l, x_k\}, y}{\texttt{minimize}} \quad t \tag{13}$$

$$\texttt{s.t.} \quad \text{(12a)},\ \text{(12b)}$$

$$A\boldsymbol{p} \leq e^{y_0}(y - y_0 + 1), \tag{13a}$$

$$1 + p_k/\sigma^2 \leq e^{x_{0k}}(x_k - x_{0k} + 1), \forall k, \tag{13b}$$

where y_0, x_{0k} are arbitrary accessible points, and $\boldsymbol{x}_0 \triangleq \{x_{0k}\}_{k=1}^K$.

We observe that problem (13) is convex since the objective function and the constraints are convex. Thus, it can be solved in polynomial time by standard solvers, e.g., CVX. Since $e^{x_0}(x - x_0 + 1) \leq e^x, \forall x_0$, the approximated problem (13) always gives a suboptimal solution of the original problem (12).

TABLE I: Iterative Algorithm to Solve (12)

1. Initialize $\boldsymbol{x}_0 \triangleq \{x_{0k}\}_{k=1}^K, y_0$, ϵ, t_{old} and `error`.
2. While `error` $> \epsilon$ do
   2.1. Solve $\boldsymbol{Q}_1(\boldsymbol{x}_0, y_0)$ in (13) to obtain the optimal values $t_\star, \boldsymbol{p}_\star, \boldsymbol{q}_\star, \boldsymbol{x}_\star, y_\star$
   2.3. Compute `error` $= |t_\star - t_{\text{old}}|$
   2.4. Update $t_{\text{old}} = t_\star, \boldsymbol{x}_0 = \boldsymbol{x}_\star, y_0 = y_\star$

We note that the optimal solution of problem (13) is largely determined by the parameters $\{x_{0k}\}_{k=1}^K, y_0$. Therefore, it is important to choose these values such that the solution of (13) is close to the optimal solution of (12). As such, we propose an iterative optimization algorithm to improve the performance of problem (13). The premise behind the proposed algorithm is to better select parameters $\{x_{0k}\}_{k=1}^K, y_0$ through iterations. The steps of the proposed algorithm are presented in Table I. The convergence of the proposed algorithm is given in the below proposition.

Proposition 1: The objective function of problem $\boldsymbol{Q}_1(\boldsymbol{x}_0 \triangleq \{x_{0k}\}_{k=1}^K, y_0)$ in (13) solved by the iterative algorithm in Table I decreases by iterations.

The proof of Proposition 1 is given in Appendix A. Although Proposition 1 does not prove the optimality of the approximated problem (13), it justifies the convergence of the proposed iterative optimal algorithm.

IV. Delivery time minimization under MMSE design

Despite the low computational complexity, the ZF-based design might result in a poor performance in some weak channel conditions. To deal with such situations, we propose an optimal power control based on the MMSE beamforming. The precoding vector in this case is given in (7). Denote $\beta_{ki} = |\boldsymbol{h}_k \check{\boldsymbol{h}}_i|^2, \forall i,k$ as the interference factor caused to user k from user i' beamforming vector, and let $\bar{\beta}_k = \| \check{\boldsymbol{h}}_k \|^2$. The achievable information of the access link for user k under the MMSE design is

$$R_{MSE,k} = W \log_2 \Big(1 + \frac{\beta_{kk} p_k}{\sum_{i \neq k} \beta_{ki} p_i + \sigma^2}\Big), \quad (14)$$

and the total transmit power at the EN is $P_{EN} = \sum_{k=1}^K \bar{\beta}_k p_k$.

The minimization problem of the largest delivery time under the MMSE design is stated as follows:

$$\begin{aligned} \underset{\{p_k,q_l\}}{\texttt{minimize}} \quad & \max\Big(\frac{Q}{R_{MSE,1}}, \ldots, \frac{Q}{R_{MSE,K}}\Big), && (15)\\ \texttt{s.t.} \quad & C_k \geq \bar{\mu}_k R_{MSE,k}, \forall k && (15a)\\ & \sum\nolimits_{k=1}^K \bar{\beta}_k p_k \leq P_\Sigma^{EN};\ \sum\nolimits_{l=1}^{\bar{L}} q_l \leq P_\Sigma^{BS}. && (15b) \end{aligned}$$

By using (14) and introducing a new variable t, problem (15) can be reformulated as follows:

$$\begin{aligned} \underset{t,\{p_k,q_l\}}{\texttt{minimize}} \quad & t && (16)\\ \texttt{s.t.} \quad & \log\Big(1 + \frac{\beta_{kk} p_k}{\sum_{i \neq k} \beta_{ki} p_i + \sigma^2}\Big) \geq \frac{Q \log(2)}{tW}, \forall k && (16a)\\ & \rho_k \sum\nolimits_{l=1}^{\bar{L}} \log\Big(1 + \frac{\lambda_l q_l}{\eta \sum_{k=1}^K \bar{\beta}_k p_k + \sigma^2}\Big)\\ & \quad \geq \bar{\mu}_k \log\Big(1 + \frac{\beta_{kk} p_k}{\sum_{i \neq k} \beta_{ki} p_i + \sigma^2}\Big), \forall k && (16b)\\ & \sum\nolimits_{k=1}^K \bar{\beta}_k p_k \leq P_\Sigma^{EN};\ \sum\nolimits_{l=1}^{\bar{L}} q_l \leq P_\Sigma^{BS}. && (16c) \end{aligned}$$

Next, we define following parameters: $\boldsymbol{\beta} = [\bar{\beta}_1, \ldots, \bar{\beta}_K]$, $B_k = [\beta_{k1}, \ldots, \beta_{k(k-1)}, 0, \beta_{k(k+1)}, \ldots, \beta_{kK}]$, and $A_k = [\beta_{k1}, \beta_{k2}, \ldots, \beta_{kK}]$. The problem 16 is equivalent to following problem:

$$\begin{aligned} \underset{t,\{p_k,q_l\}}{\texttt{min}} \quad & t && (17)\\ \texttt{s.t.} \quad & \log(A_k \boldsymbol{p} + \sigma^2) \geq \frac{Q \log(2)}{tW} + \log(B_k \boldsymbol{p} + \sigma^2), \forall k && (17a)\\ & \rho_k \sum\nolimits_{l=1}^{\bar{L}} \log(\eta \boldsymbol{\beta} \boldsymbol{p} + \lambda_l q_l + \sigma^2) + \bar{\mu}_k \log(B_k \boldsymbol{p} + \sigma^2)\\ & \quad \geq \bar{\mu}_k \log(A_k \boldsymbol{p} + \sigma^2) + \rho_k \bar{L} \log(\eta \boldsymbol{\beta} \boldsymbol{p} + \sigma^2), \forall k && (17b)\\ & \boldsymbol{\beta} \boldsymbol{p} \leq P_\Sigma^{EN};\ \sum\nolimits_{l=1}^{\bar{L}} q_l \leq P_\Sigma^{BS}, && (17c) \end{aligned}$$

where $\boldsymbol{p} \triangleq [p_1, \ldots, p_K]^T$.

We observe that problem 17 is non-convex due to the constraints (17a) and (17b). In order to leverage the non-convexity of these constraints, we introduce arbitrary positive variables $\{x_k, y_k\}_{k=1}^K$ and z, and reformulate the problem (17) as follows:

$$\begin{aligned} \underset{t,\{p_k,q_l,x_k,y_k\},z}{\texttt{min}} \quad & t && (18)\\ \texttt{s.t.} \quad & \text{(17c)};\ \log(A_k \boldsymbol{p} + \sigma^2) \geq \frac{Q \log(2)}{tW} + y_k, \forall k && (18a)\\ & \rho_k \sum\nolimits_{l=1}^{\bar{L}} \log(\eta \boldsymbol{\beta} \boldsymbol{p} + \lambda_l q_l + \sigma^2) + \bar{\mu}_k \log(B_k \boldsymbol{p} + \sigma^2)\\ & \quad \geq \bar{\mu}_k x_k + \rho_k \bar{L} z, \forall k && (18b)\\ & A_k \boldsymbol{p} + \sigma^2 \leq e^{x_k}, B_k \boldsymbol{p} + \sigma^2 \leq e^{y_k}, \forall k && (18c)\\ & \eta \boldsymbol{\beta} \boldsymbol{p} + \sigma^2 \leq e^z. && (18d) \end{aligned}$$

Although the constraints 17a and (17b) have been transformed into convex expressions, the new constraints (18c) and (18d) make problem (18) difficult to be solved optimally. Instead, we seek for a sub-optimal solution by using the inner approximations of these constraints. Similar to the previous section, we employ the first-order approximation of the exponential function in constraints (18c), (18d). In particular, let $\boldsymbol{x}_0 \triangleq \{x_{0k}\}_{k=1}^K, \boldsymbol{y}_0 \triangleq \{y_{0k}\}_{k=1}^K, z_0$ be arbitrary accessible points, we can approximate problem 18 as follows:

$$\begin{aligned} \boldsymbol{Q}_2(\boldsymbol{x}_0, \boldsymbol{y}_0, z_0): \underset{t,\{p_k,q_l,x_k,y_k\},z}{\texttt{min}} \quad & t && (19)\\ \texttt{s.t.} \quad & 18a,\ 18b\\ & A_k \boldsymbol{p} + \sigma^2 \leq e^{x_{0k}}(x_k - x_{0k} + 1), \forall k && (19a)\\ & B_k \boldsymbol{p} + \sigma^2 \leq e^{y_{0k}}(y_k - y_{0k} + 1), \forall k, && (19b)\\ & \eta \boldsymbol{\beta} \boldsymbol{p} + \sigma^2 \leq e^{z_0}(z - z_0 + 1). && (19c) \end{aligned}$$

For a known feasible set $\{x_{0k}, y_{0k}\}_{k=1}^K, z_0$, it is straightforward to verify the convexity of problem (19), since the objective function and the constraints are convex. Therefore, it can be solved in an efficient manner by standard solvers, e.g., CVX. Because $e^{\bar{y}_k}(y_k - \bar{y}_k + 1) \leq e^{y_k}, \forall \bar{y}_k$, the resorted problem (19) gives a suboptimal solution of problem (18).

It is important to note that the optimal solution of problem (19) relies on parameters $\{x_{0k}, y_{0k}\}_{k=1}^K$ and z_0. This raises a question that how to choose the values $\{x_{0k}, y_{0k}\}_{k=1}^K$ and z_0 such (19) gives a solution as close as to the optimal solution of (18). To achieve this goal, we propose an iterative optimization algorithm to improve the performance of problem (19), whose steps are listed in Table II.

TABLE II: ITERATIVE ALGORITHM TO SOLVE (18)

1. Initialize $\boldsymbol{x}_0 \triangleq \{x_{0k}\}_{k=1}^K, \boldsymbol{y}_0 \triangleq \{y_{0k}\}_{k=1}^K, z_0, \epsilon, t_{\text{old}}$ and `error`.
2. While `error` $> \epsilon$ do
   2.1. Solve $\boldsymbol{Q}_2(\boldsymbol{x}_0, \boldsymbol{y}_0, z_0)$ in (19) to obtain the optimal values $t_\star, \boldsymbol{p}_\star, \boldsymbol{q}_\star, \boldsymbol{x}_\star, \boldsymbol{y}_\star, z_\star$
   2.3. Compute `error` $= |t_\star - t_{\text{old}}|$
   2.4. Update $t_{\text{old}} = t_\star, \boldsymbol{x}_0 = \boldsymbol{x}_\star, \boldsymbol{y}_0 = \boldsymbol{y}_\star, z_0 = z_\star$.

Proposition 2: The objective function of problem $\boldsymbol{Q}_2(\boldsymbol{x}_0, \boldsymbol{y}_0, z_0)$ in (19) solved by the iterative algorithm in Table II decreases by iterations.

The proof of Proposition 2 is omitted due to the space limitation, but can be found by using similar arguments as in Proposition 1. It is evident from Proposition 2 that the proposed optimization algorithm closes the gap between the approximated problem and the original problem as the number of iterations increases.

V. PERFORMANCE EVALUATION

This section presents numerical results to demonstrate the effectiveness of our proposed optimization algorithms. The wireless channels are subject to Rayleigh fading. The pathloss on the backhaul and access channels are equal to $\sigma_{\text{bh}}^2 = -60$dB and $\sigma_{\text{ac}}^2 = -50$dB, respectively. Otherwise mentioned, the self-interference cancellation efficiency is equal to $\eta = -70$dB [13]. Other parameters are as follows: $L = N = K = 4$, $\sigma^2 = -80$ dBW, $F = 100$, $Q = 100$Mb, and $W = 10$MHz. The simulation results are calculated based on 10000 random requests over 100 channel realizations. The user requests are assumed to follow the Zipf distribution with the skewness factor $\xi = 0.8$. The Zipf-based caching policy is used, in which the most M popular files are prefetched in the EN's cache. The proposed cache-aided FD scheme is compared with the conventional HD counterpart, in which the backhaul and access transmission occur in two consecutive time slots. Therefore, the total delivery time in the HD mode is the summation of the delivery time on the backhaul link and on the access link. The delivery time of the HD mode is computed by the standard max-min design [3].

Fig. 2 plots the delivery time as a function of the WAP's transmit power, P_Σ^{BS}, with $M = 0.3F$ and $P_\Sigma^{EN} = 5$W. Two linear designs, i.e., ZF and MMSE, are shown for both FD and HD schemes. It is observed from the figure that the cache-aided FD significantly reduces the delivery time compared with the half-duplex system. At the WAP's transmit power equal to 5W, a reduction of 25% is obtained by the FD scheme with both the precoding designs. Compared with the ZF, the MMSE design obtains a 10% less in the delivery time in the observed WAP's power values. This is because the MMSE performs power allocation more effectively than the ZF in some weak conditions when the channel matrix is low rank. It is also observed that large values of P_Σ^{BS} will have less impacts on the delivery time. In this case, increasing the WAP's transmit power does not lead to zero delivery time, since it is limited by the access link for a finite P_Σ^{EN}.

Fig. 3 presents the average delivery time versus the normalized cache size, the ratio between the cache size M and the library size F, i.e., $\frac{M}{F}$. Larger cache size values result in smaller delivery times in all schemes. The benefit of caching can be also interpreted as a means of trading memory for power: the delivery time with a large transmit power ($P_\Sigma^{BS} = 10$W, $M = 0.3F$ in Fig. 2) can also be achieved with a smaller transmit power and a larger cache size ($P_\Sigma^{BS} = 1$W, $M = 0.7F$ in Fig. 3). Furthermore, the relative gain of the FD system over the HD scheme diminishes as the cache size increases. In such situations, it is highly probable that the requested file is already available at the EN's cache, thus there is less traffic on the backhaul. Note that having all the files cached does not result in zero delivery time due to the access link bottle neck.

Fig. 4 plots the delivery time versus the self-interference cancellation efficiency η. Obviously, the delivery time of the HD system is independent from the cancellation efficiency since there is not self interference in this transmission mode. It

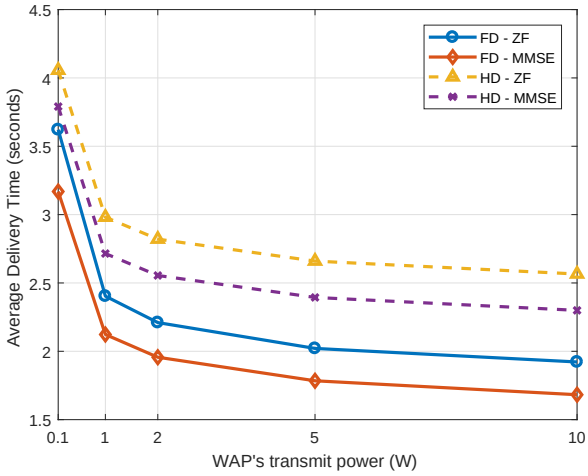

Fig. 2: Average delivery time of the cache-aided FD compared with the HD scheme v.s. the WAP's transmit power. $M = 0.3F$, $P_\Sigma^{EN} = 5$ W.

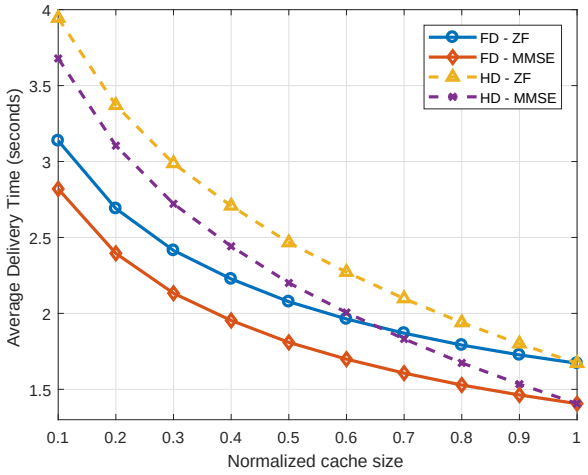

Fig. 3: Average delivery time v.s. the normalized cache size $\frac{M}{F}$. $P_\Sigma^{BS} = 1$W and $P_\Sigma^{EN} = 5$W.

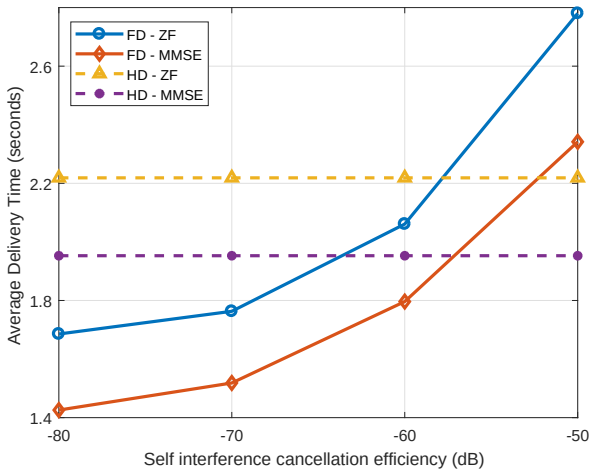

Fig. 4: Average delivery time v.s. the self-interference cancellation efficiency η. $M = 0.3F$, $P_{\Sigma}^{BS} = 10$W, and $P_{\Sigma}^{EN} = 5$W.

is shown that the FD system outperforms the HD mode in the small values of η. When the performance of the interference cancellation degrades, there is a crossing point between the FD and HD curves since the FD mode is limited by the residual interference. This result provides a guideline to determine the transmission mode when designing the cache-aided system.

VI. Conclusion

In this paper, we have investigated the performance of a cache-aided full-duplex system via delivery time analysis and optimization. Two optimization problems are formulated to minimize the average delivery time under the two linear zero-forcing and minimum mean square error precoding designs. To cope with the non-convexity of the formulated problems, we proposed two iterative optimization algorithms based on the inner approximation method. We demonstrate via numerical results the effectiveness of the cache-aided full-duplex system over the half-duplex counterpart.

The outcome of this work proves the benefits of the considered cache-aided FD system and motivates future study of cache-aided FD networks. One potential subject is the investigation on general (non-linear) precoding design, which requires the optimization of both direction and power of the beamforming vectors.

Acknowledgement

This work is supported by the Luxembourg National Research Fund under the project FNR CORE ProCAST and Vietnam National University, Hanoi, under Project No. QG.18.39.

Appendix A
Proof of Proposition 1

Denote $\left(t_{\star}^{(i)}, \boldsymbol{p}_{\star}^{(i)}, \boldsymbol{q}_{\star}^{(i)}, \boldsymbol{x}_{\star}^{(i)}, y_{\star}^{(i)}\right)$ as the optimal solution of $\boldsymbol{Q}(\boldsymbol{x}_0^{(i)}, y_0^{(i)})$ at iteration i. We will show that if $x_{\star k}^{(i)} < x_{0k}^{(i)}, \forall k$ and $y_{\star}^{(i)} > y_0^{(i)}$, then by using $x_{0k}^{(i+1)} = x_{\star k}^{(i)}, y_0^{(i+1)} = y_{\star}^{(i)}$ in the $(i+1)$-th iteration, we will have $t_{\star}^{(i+1)} < t_{\star}^{(i)}$. Indeed, by choosing a relatively large initial value $\{x_0^{(1)}\}_{k=1}^{K}$ and small value $y_0^{(1)}$, we always have $x_{\star k}^{(1)} < x_{0k}^{(1)}, \forall k$ and $y_{\star}^{(1)} < y_0^{(1)}$.

Denote $f_1(x; a) = e^a(x - a + 1)$ as the first order approximation of the e^x function at a. By using $\boldsymbol{x}_{\star}^{(i)}$ at the $(i+1)$-th iteration, we have $x_{0k}^{(i+1)} = x_{\star k}^{(i)}, \forall k$. Therefore, $f_1(x; x_{\star k}^{(i)})$ is used in the right-hand side of constraint (13b). Consider a candidate $\boldsymbol{x}^{(i+1)} = \{x_1^{(i+1)}, \ldots, x_K^{(i+1)}\}$ with $x_k^{(i+1)} \in (\hat{x}_k, x_{\star k}^{(i)})$, where $\hat{x}_k = x_{\star k}^{(i)} - 1 + e^{x_{0k}^{(i)} - x_{\star k}^{(i)}}(x_{\star k}^{(i)} - x_{0k}^{(i)} + 1)$. It is evident that $x_k^{(i+1)} < x_{\star k}^{(i)}$ and $f_1(x_k^{(i+1)}; x_{\star k}^{(i)}) > f_1(x_{\star k}^{(i)}; x_{0k}^{(i)}), \forall k$. In addition, consider a candidate $y^{(i+1)} = y_{\star}^{(i)} + \delta y$, with $\delta y <= \min_k\{\bar{\mu}_k(x_{\star k}^{(i)} - x_k^{(i+1)})\}$. Obviously, $f_1(y^{(i+1)}; y_{\star}^{(i)}) > f_1(y_{\star}^{(i)}; y_0^{(i)})$ due to the convexity of e^y function.

Because $f_1(x_k^{(i+1)}; x_{\star k}^{(i)}) > f_1(x_{\star k}^{(i)}; x_{0k}^{(i)}), \forall k$ and $f_1(y^{(i+1)}; y_{\star}^{(i)}) > f_1(y_{\star}^{(i)}; y_0^{(i)})$, the strictly inequality holds in constraints (13a) and (13b). Thus, there exits $p_k^{(i+1)} > p_{\star k}^{(i)}$ and $t^{(i+1)} < t_{\star}^{(i)}$ which satisfies constraints (12a), (13a) and (13b). Now consider a new candidate set $\left(t^{(i+1)}, \boldsymbol{p}^{(i+1)}, \boldsymbol{q}_{\star}^{(i)}, \boldsymbol{x}^{(i+1)}, \boldsymbol{y}^{(i+1)}\right)$. This set satisfies all the constraints of problem $\boldsymbol{Q}(\boldsymbol{x}_{\star}^{(i)}, \boldsymbol{y}_{\star}^{(i)})$, and therefore is a feasible solution of the optimization problem. As the result, the optimal solution at the $(i+1)$-th iteration, $t_{\star}^{(i+1)}$, must satisfy $t_{\star}^{(i+1)} \leq t^{(i+1)} < t_{\star}^{(i)}$, which completes the proof of Proposition 1.

References

[1] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, Mar. 2010, pp. 1–9.

[2] F. Gabry, V. Bioglio, and I.Land, "On energy-efficient edge caching in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3288–3298, Dec. 2016.

[3] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Edge-caching wireless networks: Performance analysis and optimziation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2827 – 2839, Apr. 2018.

[4] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[5] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Info. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[6] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, Aug. 2016.

[7] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 3, pp. 1637–1652, Sept. 2014.

[8] M. Maso, I. Atzeni, I. Ghamnia, E. Batu, and M. Debbah, "Cache-aided full-duplex small cells," in *Proc. Int. Symp. on Modeling and Opt. in Mobile, Ad Hoc, and Wireless Netw. (WiOpt)*, May 2017, pp. 1–6.

[9] M. Naslcheraghi, M. Afshang, and H. S. Dhillon, "Modeling and performance analysis of full-duplex communications in cache-enabled D2D networks," in *IEEE Int. Conf. Commun.*, May 2018, pp. 1–6.

[10] K. T. Hemachandra, O. Ochia, and A. O. Fapojuwo, "Performance study on cache enabled full-duplex device-to-device networks," in *IEEE Wireless Commun. Netw. Conf.*, April 2018, pp. 1–6.

[11] J. Kakar, A. Alameer, A. Chaaban, A. Sezgin, and A. Paulraj, "Delivery time minimization in edge caching: Synergistic benefits of subspace alignment and zero forcing," in *Proc. IEEE Int. Conf.Commun.*, May 2018, pp. 1–6.

[12] M. E. Knox, "Single antenna full duplex communications using a common carrier," in *Proc. IEEE Wireless Microwave Techno. Conf.*, April 2012, pp. 1–6.

[13] D. Bharadia and S. Katti, "Full duplex MIMO radios," in *Proc. USENIX Conf. Netw. Sys. Design Implementation*, Berkeley, CA, USA, 2014, pp. 359–372.

[14] D. Bharadia and S. Katti, "Fastforward: Fast and constructive full duplex relays," in *Proc. ACM Conf. on SIGCOMM*, Aug. 2014, pp. 199–210.