Full-Duplex Enabled Mobile Edge Caching: From Distributed to Cooperative Caching

Thang X. Vu, *Member, IEEE*, Symeon Chatzinotas, *Senior Member, IEEE*, Björn Ottersten, *Fellow, IEEE*, and Anh Vu Trinh

***Abstract*—Mobile edge caching (MEC) has received much attention as a promising technique to overcome the stringent latency and data hungry requirements in future generation wireless networks. Meanwhile, full-duplex (FD) transmission can potentially double the spectral efficiency by allowing a node to receive and transmit in the same time/frequency block simultaneously. In this paper, we investigate the delivery time performance of full-duplex enabled MEC (FD-MEC) systems, in which the users are served by distributed edge nodes (ENs), which operate in FD mode and are equipped with a limited storage memory. Firstly, we analyse the FD-MEC with different levels of cooperation among the ENs and take into account a realistic model of self-interference cancellation. Secondly, we propose a framework to minimize the system delivery time of FD-MEC under both linear and optimal precoding designs. Thirdly, to deal with the non-convexity of the formulated problems, two iterative optimization algorithms are proposed based on the inner approximation method, whose convergence is analytically guaranteed. Finally, the effectiveness of the proposed designs are demonstrated via extensive numerical results. It is shown that the cooperative scheme mitigates inter-user and self interference significantly better than the distributed scheme at an expense of inter-EN cooperation. In addition, we show that minimum mean square error (MMSE)-based precoding design achieves the best performance-complexity trade-off, compared with the zero-forcing and optimal designs.**

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

I. Introduction

Among potential enabling technologies to tackle with stringent latency and data hungry requirements in future wireless networks, mobile edge caching (MEC) has received much attention. The basic premise of MEC is to bring the content close to end users via distributed storages through out the network. Caching usually comprises a placement phase and a delivery phase. In the former, which is implemented during off-peak periods when the network resources are abundant, popular content is prefetched in the distributed caches. The latter usually occurs during peak-hours when the content requests are revealed. If the requested content is available in the edge node's local storage, it can be served directly without being sent from the core network. In this manner, MEC enables significant reduction in transmission latency and backhaul traffic thus mitigating network congestion [1]. Joint design for content caching and physical layer transmission 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 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 (EE) [2–4], [6] and content delivery time [7–9]. The role of caching in wireless device-to-device (D2D) networks is analysed in [10–12]. The performance of cache-aided wireless networks can be further improved by joint optimization of caching along with routing and resource allocation [13]. It is worth noting that these works study the caching in the half-duplex (HD) systems.

Meanwhile, full-duplex (FD) has shown great potential as the transmission technique to overcome the spectral scarceness in next generation wireless networks by allowing a node to transmit and receive in the same time/frequency resource [14–18]. The foreseen benefit of FD is, however, not without limitation. The major issue lies in the interference caused by the FD transmissions. In fact, a few FD links might result in continuous interference towards neighbouring nodes, in addition to the self interference. Fortunately, thanks to recent developments in the self-interference cancellation, FD can potentially double the spectral efficiency compared with the conventional HD counterpart [14]. The employment of FD systems with caching capability has the potential to further improve the system performance.

A. Related works

Despite that cache-aided HD has been well studied in the literature, the investigation on cache-aided FD systems is limited. The authors in [22] show that cache-aided FD small cell networks can provide cache hit enhancements compared with the HD system. In that work, by modelling the base stations and users as a coupled Poison Point Process (PPP)

Manuscript received Nov. 30, 2018; revised Apr. 9, 2019 and Aug. 27, 2019; accepted Oct. 28, 2019. The work of T. X. Vu, S. Chatzinotas and B. Ottersten is supported in by the European Research Council under project AGNOSTIC, grant R-AGR-3283 and the Luxembourg National Research Fund under project ProCAST, grant R-AGR-3415. The work of A. V. Trinh is supported by the Vietnam National University, Hanoi, under project QG.18.39. Parts of this work was presented to the conference IEEE WCNC 2019 [30]. The associate editor coordinating the review of this paper and approving it for publication was S. Ma.

T. X. Vu, S. Chatzinotas and B. Ottersten are with the Interdisciplinary Centre for Security, Reliability and Trust (SnT) – University of Luxembourg, L-1855 Luxembourg. Email: {thang.vu, symeon.chatzinotas, bjorn.ottersten}@uni.lu.

A. V. Trinh is with the Department of Electronics and Telecommunications – VNU University of Engineering and Technology, Hanoi, Vietnam. Email: vuta@vnu.edu.vn

with the edge nodes, coverage probability and successful delivery rates are analysed. The role of caching in FD D2D networks is investigated in [19], [20] via stochastic geometry analysis. By considering all possible operating modes of an arbitrary device, the success probability is derived in [19] as a function of the caching capacity and interference distribution. It is shown in [20] that allowing a hybrid deployment of FD and HD modes can potentially further improve the coverage probability in cluster-based FD networks. The authors of [21] derive closed-form expression for the successful delivery probability of a cached-aided FD system by taking into account the distribution of all wireless links. The worst case normalized delivery time (NDT) in heterogeneous networks is studied in [23] with FD relaying nodes. However, the results in [23] are based on an optimistic assumption of perfect self-interference cancellation. In practice, there always remains residual interference after the self-interference cancellation [24], [25].

B. Our contributions

In this paper, we study the performance of FD-enabled MEC (FD-MEC) systems, in which the users demand content via distributed cache-assisted edge nodes (ENs). The ENs operate in FD mode and connect to the core network via wireless backhauls. Unlike previous works on cache-assisted HD systems [3], the FD transmissions can cause significant self-interference, in addition to inter-user interference. Our goal is to minimize the system delivery time via a joint design of precoding vectors on both backhaul and access links, by taking into consideration the cached content and interference patterns. The contributions of this paper are as follows:

- Firstly, we investigate the delivery time performance of FD-MEC systems by considering a realistic model of the self-interference cancellation. Our work is fundamentally different from [19–22] in terms of performance metric and analysis method. Compared with [23], which understands the cache-aided FD system from the information-theoretic asymptotic aspect and relies on the perfect assumption of self-interference cancellation, we consider the practical interference cancellation model and focus on precoding vectors design.
- Secondly, we analyse the system via two network architectures for different levels of cooperation among the ENs, namely distributed and cooperative caching. For each architecture, an optimization problem is formulated that minimizes the system delivery time based on both linear and optimal (non-linear) precoding designs. The formulated problems optimize the precoding vectors while minimizing both inter-user and self interference.
- Thirdly, to cope with the non-convexity of the formulated problems caused by the self-interference, we propose two iterative optimization algorithms based on the inner approximation method. The convergence of the proposed iterative algorithms are analytically guaranteed.
- Finally, extensive numerical results are presented to demonstrate the effectiveness of the proposed algorithms and the benefit of the FD-MEC over the half-duplex system in certain scenarios.

The rest of this paper is organised as follows. Section II presents the system model and the caching strategies. Section III gives the signal transmission details of the two caching modes. Section IV proposes the precoding vectors design for the distributed caching scheme. Section V optimizes the precoding vectors for the cooperative caching scheme. Numerical results are shown in Section VI. Finally, Section VII provides conclusions and discussions.

Notation: $(.)^H, (.)^T$ and $(.)^{-1}$ denote the conjugate operator, transpose operator, and the inverse matrix, respectively. $\mathrm{Tr}(\boldsymbol{X})$ denotes the trace of matrix $\boldsymbol{X}$.

II. System and caching model

We consider a cache-aided FD system, in which the users are served via a number of distributed ENs, e.g., pico- and small cell base stations, as depicted in Fig. 1. The ENs operate in FD mode and connect to the core network via a wireless backhaul access point (WAP), e.g., macro base station. The users can only access data from the ENs via wireless access channels, i.e., there is no direct link between the users and the WAP. The WAP is equipped with N antennas, while the ENs and users are equipped with a single-antenna. Let K denote the number of ENs with $K \leq N$. It is assumed that each EN serves only one user at a time [22], thus the number of active users is also K[1]. The WAP is assumed to have access to a library of F contents, denoted by $\mathcal{F} = \{f_1, \ldots, f_F\}$. Without loss of generality, all content is assumed to have equal size of Q bits. To leverage the backhaul during peak-hours, the EN is equipped with a storage memory of MQ bits, where $M < F$.

We consider two network architectures depending on the level of collaboration among the ENs: i) *Distributed caching - separate access transmission* (DCST) and ii) *Cooperative caching with joint access transmission* (CCJT).

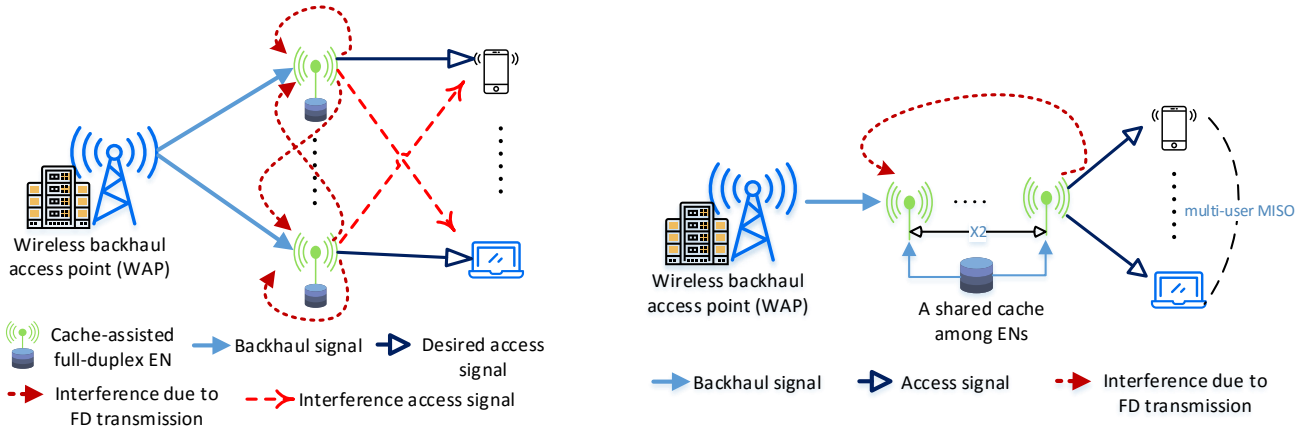
A. Distributed caching - separate transmission (DCST) mode

The DCST mode does not require any collaboration among the ENs, hence minimizing the system's signalling overhead (Fig. 1a). In this mode, each EN stores the content in its local cache, and operates independently from other ENs. The benefit of this mode is the scalability and flexibility. However, since the transmissions of the ENs are independent, severe inter-user interference can occur, resulting in a significant performance degradation. Details on this mode will be presented in the next section.

B. Cooperative caching - joint transmission (CCJT) mode

In this mode, all the ENs share a common cache and cooperatively serve the user requests (Fig. 1b). Such cooperation can be enabled via the dedicated X2 link [26]. Since the ENs jointly transmit the requested contents to the users, inter-user interference can be efficiently mitigated, thus the system performance can be largely improved compared to the DCST mode. These improvements, however, require inter-EN

[1]When the number of users is greater than K, the EN can serves its active users via, e.g., time division multiplexing. Studying such scenario is beyond the scope of this paper.



(a) Distributed architecture

(b) Cooperative architecture

Fig. 1: Block diagram of FD-MEC. In the distributed caching architecture (a), each EN has its own cache and serves its intended user separately, which imposes three types of interferences at the users and ENs's (backhaul) receiver. In the cooperative caching architecture (b), the ENs share a common cache and cooperatively serve the users via joint access transmission. In both architectures, the ENs decode the backhaul signal separately.

collaboration and additional signalling overhead. It is worth noting that although the inter-user interference can be avoided, self-interference still exits at the ENs' receivers due to the ENs' FD transmission.

Remark 1: Intuitively, these two modes serve as the two extremes of the network architecture when inter-EN cooperation is allowed. Analysing these two modes provide the lower- and upper-bound for the performance of FD-MEC systems.

C. Content popularity and caching model

We consider the most popular content popularity model, i.e., the Zipf distribution [27]. The probability for file f_n being requested is equal to

$$\nu_n = \frac{n^{-\xi}}{\sum_{m=1}^{F} m^{-\xi}}, \tag{1}$$

where ξ is the Zipf skewness factor.

This paper focuses on off-line caching delivery phase, in which the *content placement phase* is predetermined and executed during off-peak times [3–6], [10], [12]. We consider a generic caching policy (cache placement) $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_F]$, where $0 \leq \mu_n \leq 1$ denotes portions of file f_n cached at the ENs. 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 studying different caching strategies. In the most popular caching, we have $\boldsymbol{\mu}_{\text{Pop}} = [\underbrace{1, \ldots, 1}_{\times M}, 0, \ldots, 0]$.

III. Signal Transmission Model

In this section, we provide details on the signal transmission in two DCST and CCJT modes. Since each EN serves only one user, we use the user index and EN index interchangeably, e.g., EN k means the EN serving user k. When user k demands a file, it sends the requested file index d_k to its serving EN k. The EN k first checks its local cache. If (portions of) the requested content is available in the cache, it serves the user directly. Otherwise, the EN k will ask for the non-cached parts from the WAP via the wireless backhaul before serving the user.

Denote $\mathcal{U}_C$ as the set of ENs which have only some portions of the requested files in their caches, i.e., $\mathcal{U}_C = \{k \mid \mu_{d_k} < 1\}$. Let $K_C = |\mathcal{U}_C|$. Without loss of generality, *we assume the first K_C ENs cache only parts of the requested files* for ease of presentation, i.e., $\mathcal{U}_C = \{1, 2, \ldots, K_C\}$. Because each EN serves only one user, we also refer $\mathcal{U}_C$ as the set of the users served by the ENs in $\mathcal{U}_C$. This way, any EN $l \notin \mathcal{U}_C$, i.e., $K_C < l \leq K$, has the whole requested file in its cache.

The channel fading coefficients, including the path loss, of all the links are defined in Table I. Full channel state information is assumed to be known at the transmitter sides.

A. Signal transmission in DCST mode

1) Signal transmission on backhaul links: Since the ENs not in $\mathcal{U}_C$ have the whole requested files in their cache, the WAP only sends the non-cached parts of the requested files to the ENs in $\mathcal{U}_C$ via the backhaul. Let $s_{E,k}, \forall k \in \mathcal{U}_C$, denote the data symbol target to EN k from the WAP. The WAP first precodes the data before sending on the backhaul. In this paper, we consider a linear minimum mean square error (MMSE) precoding for the backhaul transmission. In this design, the

TABLE I: Channel fading coefficients, including the pathloss

Notation	Explanation.
$\boldsymbol{g}_k \in \mathbb{C}^{1\times N}$	WAP $\rightarrow$ EN k backhaul channel coefficients
$f_{kl} \in \mathbb{C}$	EN $l \rightarrow$ EN k inter-EN channel coefficients due to the FD transmission
$f_{kk} \in \mathbb{C}$	Self-interference at EN k due to the FD transmission
$h_{kl} \in \mathbb{C}$	EN $l \rightarrow$ user k access channel coefficient
$\boldsymbol{h}_k \in \mathbb{C}^{1\times K}$	$[h_{k1}, h_{k2}, \ldots, h_{kK}]$ - channel coefficients from all ENs to user k

beamforming vector for EN $k \in \mathcal{U}_C$ is given as $\boldsymbol{w}_k = \sqrt{q_k}\tilde{\boldsymbol{w}}_k$, where q_k is the power factor allocated to EN k and $\tilde{\boldsymbol{w}}_k$ is the MMSE beamforming vector that is the k-th column of the MMSE beamforming matrix $\boldsymbol{G}_C^H(\boldsymbol{G}_C\boldsymbol{G}_C^H + \sigma^2 I)^{-1}$, where $\boldsymbol{G}_C$ is the channel matrix from the WAP's antennas to the ENs in $\mathcal{U}_C$, i.e., $\boldsymbol{G}_C = [\boldsymbol{g}_1^T, \ldots, \boldsymbol{g}_{K_C}^T]^T$. The size of $\boldsymbol{G}_C$ is $K_C \times L$.

The received signal at EN $k \in \mathcal{U}_C$, $y_{E,k}$, is given as

$$y_{\mathrm{E},k} = \boldsymbol{g}_k^H \boldsymbol{w}_k s_{\mathrm{E},k} + \sum\nolimits_{k\neq l\in\mathcal{U}_C} \boldsymbol{g}_k^H \boldsymbol{w}_l s_{\mathrm{E},l} + \sqrt{p_k} f_{kk} s_{\mathrm{U},k} + \sum\nolimits_{k\neq l=1}^{K} \sqrt{p_l} f_{kl} s_{\mathrm{U},l} + n_{\mathrm{E},k}, \tag{2}$$

where $s_{U,k}$ is the transmit symbol from EN k to user k. In (2), the first term is the intended signal for EN k; the second term represents the inter-EN interference on the backhaul channels; the third term represents the self-interference at EN k due to the FD transmission; the fourth term is interference at EN k due to the FD transmission of other ENs; and $n_{E,k}$ is the Gaussian noise with zero mean and variance σ^2.

In order to decode $y_{E,k}$, EN k performs interference cancellation to mitigate the self interference, since $s_{U,k}$ is known. After the interference cancellation, there remains a residual interference with power ηp_k, where η is a Gamma distributed random variable [29] representing the self-interference cancellation efficiency with a mean $\bar{\eta}$. The common value of $\bar{\eta}$ is less than $-40dB$ depending on the hardware and interference cancellation techniques [24], [25]. We note that although the self interference can be effectively eliminated, there remain two interfering signals (the second and fourth terms in (2)). By treating interference as noise, the backhaul achievable information rate for EN $k \in \mathcal{U}_C$ is given as

$$C_{\mathtt{dist},k} = W \log\Big(1 + \frac{|\boldsymbol{g}_k^H \tilde{\boldsymbol{w}}_k|^2 q_k}{I_{\mathtt{dist},k} + \sigma^2}\Big), \tag{3}$$

where W is the channel bandwidth and $I_{\mathtt{dist},k} = \sum_{k\neq l\in\mathcal{U}_C} |\boldsymbol{g}_k^H \tilde{\boldsymbol{w}}_l|^2 q_l + \eta p_k + \sum_{k\neq i=1}^{K} p_i |f_{ki}|^2$ is the total interference at EN k.

The total transmit power at the WAP is $P_{BS} = \sum_{k\in\mathcal{U}_C} \| \tilde{\boldsymbol{w}}_k \|^2 q_k = \sum_{k=1}^{K_C} \| \tilde{\boldsymbol{w}}_k \|^2 q_k$.

2) Signal transmission on access links: In the distributed caching architecture, each EN serves its user independently. Let $s_{U,k}$ denote the signal sent from EN k to user k. The received signal at user k, $\forall k$, is given as

$$y_{U,k} = \sqrt{p_k} h_{kk} s_{U,k} + \sum\nolimits_{k\neq i=1}^{K} \sqrt{p_i} h_{ki} s_{U,i} + n_{U,k}, \tag{4}$$

where h_{ki} is defined in Table I and $n_{U,k}$ is the Gaussian noise with zero mean and variance σ^2. The second term in (4) represents the inter-user interference on the access links caused by the transmission of other ENs.

By treating interference as noise, the achievable information rate for user k (on the access links) is

$$R_{\mathtt{dist},k} = W \log\Big(1 + \frac{p_k |h_{kk}|^2}{\sum_{k\neq i=1}^{K} p_i |h_{ki}|^2 + \sigma^2}\Big). \tag{5}$$

B. Signal transmission in CCJT mode

1) Signal transmission on backhaul links: The backhaul transmission in CCJT mode is similar to the one in DCST mode. However, the access transmission in CCJT mode is different from DCST mode, since the ENs jointly serve the users. Let x_k denote the transmit signal from EN k to user k (on the access links), whose details will be presented in Sec. III-B2. The received backhaul signal at EN $k \in \mathcal{U}_C$ is given as follows:

$$y_{\mathrm{E},k} = \boldsymbol{g}_k^H \boldsymbol{w}_k s_{\mathrm{E},k} + n_{E,k} + \underbrace{\sum\nolimits_{k\neq l\in\mathcal{U}_C} \boldsymbol{g}_k^H \boldsymbol{w}_l s_{\mathrm{E},l}}_{(a)} + \underbrace{\sum\nolimits_{i=1}^{K} f_{ki} x_i}_{(b)}, \tag{6}$$

where (a) is the interference on the backhaul links, (b) is the interference due to the FD transmission of all the ENs, and $n_{E,k}$ is the Gaussian noise with zero mean and variance σ^2.

In the cooperative mode, the precoding vectors and transmitted data are shared among all the ENs. Therefore, $x_k, \forall k$ is known at every EN. In order to decode $y_{E,k}$, the EN k first performs interference cancellation on the aggregated interference (b). After self-interference cancellation, there remains a residual interference with power $\eta \sum_{l=1}^{K} |x_l|^2$, where η is the self-interference cancellation efficiency [25], [29] which is modelled as a Gamma distributed random variable with a mean $\bar{\eta}$. By treating interference as noise, the backhaul achievable information rate for EN k, $\forall k \in \mathcal{U}_C$, is given as

$$C_{\mathtt{coop},k} = W \log\Big(1 + \frac{|\boldsymbol{g}_k^H \tilde{\boldsymbol{w}}_k|^2 q_k}{I_{\mathtt{coop},k} + \sigma^2}\Big), \tag{7}$$

where $I_{\mathtt{coop},k} = \sum_{k\neq l\in\mathcal{U}_C} |\boldsymbol{g}_k^H \tilde{\boldsymbol{w}}_l|^2 q_l + \eta \sum_{l=1}^{K} |x_l|^2$.

2) Signal transmission on access links: In CCJT mode, the ENs serve all users in a cooperative manner. Therefore, the access links can be seen as a multi-user MISO channel $\boldsymbol{H} = [\boldsymbol{h}_1^T, \ldots, \boldsymbol{h}_K^T]^T$, where $\boldsymbol{h}_k = [h_{k1}, \ldots, h_{kK}]$ is the channel fading vector from all the ENs to user k. The ENs first jointly precode the data before transmitting to the users.

Let $\boldsymbol{v}_k \in \mathbb{C}^{K\times 1}$ denote a generic precoding vector for user k, and $v_k[l]$ denote the l-th element of $\boldsymbol{v}_k$. The transmit signal at EN k is $x_k = \sum_{l=1}^{K} v_l[k] s_{U,l}$, where $s_{U,l}$ is the data symbol dedicated to user l.

The received signal at user k in the CCJT mode is

$$y_{U,k} = \sum_{l=1}^{K} h_{kl} x_l + n_{U,k} = \sum_{l=1}^{K} h_{kl} \sum_{i=1}^{K} v_i[l] s_{U,i} + n_{U,k} = \boldsymbol{h}_k \boldsymbol{v}_k s_{U,k} + \sum\nolimits_{i\neq k} \boldsymbol{h}_k \boldsymbol{v}_i s_{U,i} + n_{U,k}, \tag{8}$$

where the first term is the desired signal, the second term is the aggregated inter-user interference, and $n_{U,k}$ is the Gaussian noise with zero mean and variance σ^2.

By treating interference as noise, the achievable information rate (on the access link) for user k is

$$R_{\mathtt{coop},k} = W \log\Big(1 + \frac{|\boldsymbol{h}_k \boldsymbol{v}_k|^2}{\sum_{l\neq k} |\boldsymbol{h}_k \boldsymbol{v}_l|^2 + \sigma^2}\Big). \tag{9}$$

Remark 2: Although the DCST and CCJT employ different transmission policies on the access channels, they use the same

MMSE design for the backhaul to moderate the overhead signal among the ENs. In both cases, the ENs decodes the backhaul signal individually.

Remark 3: In the cooperative architecture, the ENs employ different precoding designs on the access links, e.g., ZF, MMSE, and Optimal design, which results in different achievable rates on the access links, $R_{\texttt{coop},k}$, and different ENs' transmit powers, $\sum_{k=1}^{K} \| \boldsymbol{v}_k \|^2$, which eventually affects the backhaul information rate.

IV. Delivery time minimization in Distributed caching mode

In this section, we propose a power allocation to minimize the delivery time in DCST mode. For an EN which has the whole requested files in its cache, e.g., EN $k \notin \mathcal{U}_C$, the delivery time for this EN to serve its user is $t_k = \frac{Q}{R_k}, \forall k \notin \mathcal{U}_C$, i.e., $K_C < k \leq K$, where Q is the file size.

In order to serve a user $k \in \mathcal{U}_C$, the EN k will receive the non-cached parts from the WAP while serving its user in the FD mode. Assuming that the FastForward FD transmission is employed by the ENs [28], the delivery time for the user k, $\forall k \in \mathcal{U}_C$, is $t_k = \frac{Q}{R_k}$ subjected to a constraint that the EN's buffer is not empty. Because $\mu_{d_k} Q$ bits of the requested file is already available at the EN k's cache, this condition reads $C_{\texttt{dist},k}\tau + \mu_{d_k} Q \geq R_k \tau, \forall \tau \in [0, t_k]$. Consider all possible values of $\tau \in [0, t_k]$, this constraint becomes $C_{\texttt{dist},k} \geq \bar{\mu}_k R_k, \forall k \in \mathcal{U}_C$, where $\bar{\mu}_k \triangleq 1 - \mu_{d_k}$ represents the volume of the non-cached parts of the requested file f_{d_k}.

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

$$\underset{\{p_k\}_{k=1}^{K}, \{q_k\}_{k=1}^{K_C}}{\texttt{minimize}} \quad \max\Big(\frac{Q}{R_1}, \ldots, \frac{Q}{R_K}\Big), \tag{10}$$

$$\texttt{s.t. } C_{\texttt{dist},k} \geq \bar{\mu}_k R_k, \forall k \in \mathcal{U}_C \tag{10a}$$

$$\sum_{k \in \mathcal{U}_C} \| \tilde{\boldsymbol{w}}_k \|^2 \, q_k \leq P_{BS}; p_k \leq P_{EN}, \forall k, \tag{10b}$$

where the first constraint is to guarantee the EN's cache is not empty, P_{BS} and P_{EN} are the maximum transmit power at the WAP and the ENs, respectively. Although the objective function of problem (10) can be transformed into the max-min rate problem, the key challenge lies in the non-convexity of constraint (10a).

For ease of presentation, let $\boldsymbol{p} = [p_1, \ldots, p_K, 1]^T$ and $\boldsymbol{q} = [q_1, \ldots, q_{K_C}, 1]^T$ denote the compound power variables. In addition, we define following parameters:

$$\begin{aligned}
A_{1k} =& [|\boldsymbol{g}_k^H \tilde{\boldsymbol{g}}_1|^2, \ldots, |\boldsymbol{g}_k^H \tilde{\boldsymbol{g}}_{K_C}|^2, \sigma^2] \\
A_{2k} =& \\
& [|\boldsymbol{g}_k^H \tilde{\boldsymbol{g}}_1|^2, \ldots, |\boldsymbol{g}_k^H \tilde{\boldsymbol{g}}_{k-1}|^2, 0, |\boldsymbol{g}_k^H \tilde{\boldsymbol{g}}_{k+1}|^2, \ldots, |\boldsymbol{g}_k^H \tilde{\boldsymbol{g}}_{K_C}|^2, \sigma^2] \\
B_{1k} =& [|h_{k1}|^2, \ldots, |h_{kN}|^2, \sigma^2] \\
B_{2k} =& [|h_{k1}|^2, \ldots, |h_{k(k-1)}|^2, 0, |h_{k(k+1)}|^2, \ldots, |h_{kN}|^2, \sigma^2] \\
D_k =& [|f_{k1}|^2, \ldots, |f_{k(k-1)}|^2, \eta, |f_{k(k+1)}|^2, \ldots, |f_{kN}|^2, 0] \\
\boldsymbol{\lambda} =& [\| \tilde{\boldsymbol{g}}_1 \|^2, \ldots, \| \tilde{\boldsymbol{g}}_{K_C} \|^2, 0].
\end{aligned}$$

From (3) and (5), we can write then backhaul and access information rate as follows:

$$\begin{aligned}
C_{\texttt{dist},k} =& W \log_2\Big(\frac{D_k \boldsymbol{p} + A_{1k}\boldsymbol{q}}{D_k \boldsymbol{p} + A_{2k}\boldsymbol{q}}\Big) = \\
& W \log_2(D_k \boldsymbol{p} + A_{1k}\boldsymbol{q}) - W \log_2(D_k \boldsymbol{p} + A_{2k}\boldsymbol{q}), \forall k \in \mathcal{U}_C
\end{aligned} \tag{11}$$

$$\begin{aligned}
R_{\texttt{dist},k} =& W \log_2\Big(\frac{B_{1k}\boldsymbol{p}}{B_{2k}\boldsymbol{p}}\Big) \\
=& W \log_2(B_{1k}\boldsymbol{p}) - W \log_2(B_{2k}\boldsymbol{p}), \forall k.
\end{aligned} \tag{12}$$

By introducing a positive variable t, and using (11) and (12), the problem (10) is equivalent to the following problem:

$$\underset{t, \boldsymbol{p}, \boldsymbol{q}}{\texttt{minimize}} \quad t \tag{13}$$

$$\texttt{s.t.} \quad \log(B_{1k}\boldsymbol{p}) \geq \frac{Q \log(2)}{Wt} + \log(B_{2k}\boldsymbol{p}), \forall k \tag{13a}$$

$$\begin{aligned}
& \log(D_k \boldsymbol{p} + A_{1k}\boldsymbol{q}) + \bar{\mu}_k \log(B_{2k}\boldsymbol{p}) \\
& \geq \bar{\mu}_k \log(B_{1k}\boldsymbol{p}) + \log(D_k \boldsymbol{p} + A_{2k}\boldsymbol{q}), \forall k \in \mathcal{U}_C
\end{aligned} \tag{13b}$$

$$\boldsymbol{\lambda}\boldsymbol{q} \leq P_{BS}; \; p_k \leq P_{EN}, \forall k, \tag{13c}$$

where the new constraint (13a) results from $t \geq \frac{Q}{R_{\texttt{dist},k}}, \forall k$.

It is observed that problem (13) is non-convex since the first two constraints are non-affine. To overcome this difficulty, we will represent these constraints in a convex expression via arbitrary intermediate variables $\{x_k, z_k\}_{k=1}^{K_C}$, $\{y_k\}_{k=1}^{K}$, and reformulate problem (13) as follows:

$$\underset{t, \boldsymbol{p}, \boldsymbol{q}, \{y_k\}_{k=1}^{K}, \{x_k, z_k\}_{k=1}^{K_C}}{\texttt{minimize}} \quad t \tag{14}$$

$$\texttt{s.t.} \quad \log(B_{1k}\boldsymbol{p}) \geq \frac{Q \log(2)}{Wt} + y_k, \forall k \tag{14a}$$

$$\begin{aligned}
& \log(D_k \boldsymbol{p} + A_{1k}\boldsymbol{q}) + \bar{\mu}_k \log(B_{2k}\boldsymbol{p}) \\
& \geq \bar{\mu}_k x_k + z_k, \; 1 \leq k \leq K_C
\end{aligned} \tag{14b}$$

$$B_{1k}\boldsymbol{p} \leq e^{x_k}, \; 1 \leq k \leq K_C \tag{14c}$$

$$B_{2k}\boldsymbol{p} \leq e^{y_k}, \forall k \tag{14d}$$

$$D_k \boldsymbol{p} + A_{2k}\boldsymbol{q} \leq e^{z_k}, \; 1 \leq k \leq K_C \tag{14e}$$

$$\boldsymbol{\lambda}\boldsymbol{q} \leq P_{BS}; \; p_k \leq P_{EN}, \forall k. \tag{14f}$$

Although constraints (14a) and (14b) are now convex, solving problem (14) is still challenging since constraints (14c) - (14e) are unbounded. Fortunately, because the function e^x is convex, we can employ the inner approximation method, which replaces constraints (14c) - (14e) by using the first-order approximation of the exponential function, i.e., $e^x \simeq e^{x_0}(x - x_0 + 1)$, where x_0 is any accessible point. The approximated problem is formulated, for a given set of accessible points $\boldsymbol{x}_0 \triangleq \{x_{0k}\}_{k=1}^{K_C}, \boldsymbol{y}_0 \triangleq \{y_{0k}\}_{k=1}^{K}, \boldsymbol{z}_0 \triangleq \{z_{0k}\}_{k=1}^{K_C}$, as follows:

$$\boldsymbol{Q}_1(\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0): \underset{t, \boldsymbol{p}, \boldsymbol{q}, \{x_k, z_k\}_{k=1}^{K_C}, \{y_k\}_{k=1}^{K}}{\texttt{minimize}} \quad t \tag{15}$$

$$\texttt{s.t.} \quad \text{(14a), (14b), (14f)}$$

$$B_{1k}\boldsymbol{p} \leq e^{x_{0k}}(x_k - x_{0k} + 1), 1 \leq k \leq K_C \tag{15a}$$

$$B_{2k}\boldsymbol{p} \leq e^{y_{0k}}(y_k - y_{0k} + 1), \forall k \tag{15b}$$

$$A_{2k}\boldsymbol{q} + D_k \boldsymbol{p} \leq e^{z_{0k}}(z_k - z_{0k} + 1), 1 \leq k \leq K_C. \tag{15c}$$

It is straightforward to verify that, for a given set of $\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0$, problem (15) is convex since the objective function and the constraints are convex. Thus, it can be solved in

TABLE II: ITERATIVE ALGORITHM TO SOLVE (14)

1. Initialize $\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0$, ϵ, t_{old} and $\texttt{error}$.
2. While $\texttt{error} > \epsilon$ do
   2.1. Solve $\boldsymbol{Q}(\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0)$ in (15) to obtain the optimal values $t_\star, \boldsymbol{p}_\star, \boldsymbol{q}_\star, \boldsymbol{x}_\star, \boldsymbol{y}_\star, \boldsymbol{z}_\star$
   2.3. Compute $\texttt{error} = |t_\star - t_{\text{old}}|$
   2.4. Update $t_{\text{old}} = t_\star, \boldsymbol{x}_0 = \boldsymbol{x}_\star, \boldsymbol{y}_0 = \boldsymbol{y}_\star, \boldsymbol{z}_0 = \boldsymbol{z}_\star$

an efficient manner by standard solvers, e.g., CVX. Since $e^{x_0}(x - x_0 + 1) \leq e^x, \forall x_0$, the approximated problem (15) always gives a suboptimal solution of the original problem (14).

We note that the optimal solution of problem (15) is largely determined by the parameters $\{x_{0k}, z_{0k}\}_{k=1}^{K_C}, \{y_{0k}\}_{k=1}^{K}$. Therefore, it is important to choose proper values $\{x_{0k}, y_{0k}, z_{0k}\}$ such that the solution of (15) approaches quickly the optimal solution of (14). As such, we propose an iterative optimization algorithm to improve the performance of problem (15). The premise behind the proposed algorithm is to better select the parameters $\{x_{0k}, z_{0k}\}_{k=1}^{K_C}, \{y_{0k}\}_{k=1}^{K}$ through iterations. The details of the proposed algorithm are presented in Table II.

The convergence of the proposed iterative algorithm is guaranteed in the proposition below.

Proposition 1: The objective function of problem $\boldsymbol{Q}_1(\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0)$ in (15) solved by the iterative algorithm in Table II decreases by iterations.

Proof: See Appendix A. ∎

Although Proposition 1 does not guarantee the optimality of the approximated problem, it provides justification for the proposed iterative optimization algorithm.

V. DELIVERY TIME MINIMIZATION IN COOPERATIVE CACHING MODE

In this section, we minimize the delivery time under the CCJT mode. Intuitively, the cooperative caching mode not only reduces inter-user interference on the access links, but also improves the self-interference cancellation at the ENs since the ENs' transmit signals are shared among the ENs.

We consider three precoding designs for the access links: ZF, MMSE and optimal design which jointly optimizes the direction and magnitude of the precoding vectors. We note that the WAP employs the same backhaul precoding design as in Section IV.

A. Delivery time minimization under ZF design

The precoding vector under the ZF design is given as $\boldsymbol{v}_k = \sqrt{p_k}\breve{\boldsymbol{h}}_k$, where p_k is the power factor allocated for user k and $\breve{\boldsymbol{h}}_k$ is the ZF beamforming vector, which is the k-th column of the ZF precoding matrix $\boldsymbol{H}^H(\boldsymbol{H}\boldsymbol{H}^H)^{-1}$. In this design, the inter-user interference (on the access links) is fully cancelled, i.e., $|\boldsymbol{h}_k^H \breve{\boldsymbol{h}}_i| = \delta_{ki}, \forall k, i$. From (3) and (5) we have the backhaul and access rates under the ZF design as follows:

$$C_{\texttt{coop},k}^{ZF} = W\log_2\Big(1+\frac{|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_k|^2 q_k}{\sum\limits_{k\neq l\in\mathcal{U}_C}|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_l|^2 q_l+\eta\sum\limits_{i=1}^{K}\|\breve{\boldsymbol{h}}_i\|^2 p_i+\sigma^2}\Big), \forall k \in \mathcal{U}_C$$

$$R_{\texttt{coop},k}^{ZF} = W\log_2\Big(1+\frac{p_k}{\sigma^2}\Big), \forall k.$$

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

$$\underset{\{p_k\}_{k=1}^K, \{q_k\}_{k=1}^{K_C}}{\texttt{minimize}} \quad \max\Big(\frac{Q}{R_{\texttt{coop},1}^{ZF}}, \ldots, \frac{Q}{R_{\texttt{coop},K}^{ZF}}\Big), \tag{16}$$

$$\texttt{s.t.} \quad C_{\texttt{coop},k}^{ZF} \geq \bar{\mu}_k R_{\texttt{coop},k}^{ZF}, \forall k \in \mathcal{U}_C \tag{16a}$$

$$\textstyle\sum_{k\in\mathcal{U}_C}\|\tilde{\boldsymbol{w}}_k\|^2 q_k \leq P_{BS} \tag{16b}$$

$$\textstyle\sum_{k=1}^K \|\breve{\boldsymbol{h}}_i\|^2 p_k \leq K P_{EN}, \tag{16c}$$

where the constraint (16c) benefits from power allocation among the ENs due to the ENs' joint transmission.

Denote $t = \max\Big(\frac{Q}{R_{\texttt{coop},1}^{ZF}}, \ldots, \frac{Q}{R_{\texttt{coop},K}^{ZF}}\Big)$ as a new variable. Then problem (16) is equivalent to the following problem:

$$\underset{t,\{p_k\}_{k=1}^K, \{q_k\}_{k=1}^{K_C}}{\texttt{minimize}} \quad t \tag{17}$$

$$\texttt{s.t.} \quad \log(1+\tfrac{p_k}{\sigma^2}) \geq \tfrac{Q\log(2)}{Wt}, \forall k \tag{17a}$$

$$\log\Big(1+\frac{|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_k|^2 q_k}{\sum_{k\neq l\in\mathcal{U}_C}|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_l|^2 q_l+\eta\sum\limits_{i=1}^{K}\|\breve{\boldsymbol{h}}_i\|^2 p_i+\sigma^2}\Big)$$

$$\geq \bar{\mu}_k\log(1+\tfrac{p_k}{\sigma^2}), 1 \leq k \leq K_C \tag{17b}$$

$$(16b), (16c).$$

For ease of presentation, let us define parameters $A_{1k}, A_{2k}, \boldsymbol{\lambda}$ as in Sec. IV, and $\boldsymbol{\alpha} \triangleq [\|\breve{\boldsymbol{h}}_1\|^2, \ldots, \|\breve{\boldsymbol{h}}_K\|^2]$.

Furthermore, we use the compound notation for the powers $\boldsymbol{p} = [p_1, \ldots, p_K]^T$ and $\boldsymbol{q} = [q_1, \ldots, q_{K_C}, 1]^T$.

Then the problem (17) can be reformulated as follows:

$$\underset{t,\boldsymbol{p},\boldsymbol{q}}{\texttt{minimize}} \quad t \tag{18}$$

$$\texttt{s.t.} \quad \log(1+\frac{p_k}{\sigma^2}) \geq \frac{Q\log(2)}{tW}, \forall k \tag{18a}$$

$$\log(A_{1k}\boldsymbol{q}+\eta\boldsymbol{\alpha}\boldsymbol{p}) \geq \bar{\mu}_k\log(1+\frac{p_k}{\sigma^2}) + \log(\eta\boldsymbol{\alpha}\boldsymbol{p}+A_{2k}\boldsymbol{q}), 1 \leq k \leq K_C \tag{18b}$$

$$\boldsymbol{\alpha}\boldsymbol{p} \leq KP_{EN};\ \boldsymbol{\gamma}\boldsymbol{q} \leq P_{BS}, \tag{18c}$$

It is observed that problem (18) is non-convex since the first two constraints are non-affine. By introducing arbitrary variables $\{x_k, y_k\}_{k=1}^{K_C}$, we can reformulate problem (18) as

$$\underset{t,\boldsymbol{p},\boldsymbol{q},\{x_k,y_k\}_{k=1}^{K_C}}{\texttt{minimize}} \quad t \tag{19}$$

$$\texttt{s.t.} \quad (18a), (18c)$$

$$\log(A_{1k}\boldsymbol{q}+\eta\boldsymbol{\alpha}\boldsymbol{p}) \geq \bar{\mu}_k x_k + y_k, 1 \leq k \leq K_C \tag{19a}$$

$$1+\frac{p_k}{\sigma^2} \leq e^{x_k}, 1 \leq k \leq K_C \tag{19b}$$

$$\eta\boldsymbol{\alpha}\boldsymbol{p}+A_{2k}\boldsymbol{q} \leq e^{y_k}, 1 \leq k \leq K_C. \tag{19c}$$

It is evident that problem (19) is non-convex since the two last constraints (19b) and (19c) are unbounded. Similarly to the previous section, we employ the linear-approximation of the exponential function to approximate these two constraints. Let's x_{0k}, y_{0k} be any accessible points, the constraints (19b) and (19c) can be approximated as follows:

$$1+\frac{p_k}{\sigma^2} \leq e^{x_{0k}}(x_k - x_{0k} + 1), 1 \leq k \leq K_C \tag{19d}$$

$$\eta\boldsymbol{\alpha}\boldsymbol{p}+A_{2k}\boldsymbol{q} \leq e^{y_{0k}}(y_k - y_{0k} + 1), 1 \leq k \leq K_C. \tag{19e}$$

TABLE III: ITERATIVE ALGORITHM TO SOLVE (19)

1. Initialize $\boldsymbol{x}_0 \triangleq \{x_{0k}\}_{k=1}^{K_C}, \boldsymbol{y}_0 \triangleq \{y_{0k}\}_{k=1}^{K_C}$, ϵ, t_{old} and $\texttt{error}$.
2. While $\texttt{error} > \epsilon$ do
   2.1. Solve $\boldsymbol{Q}_2(\boldsymbol{x}_0, \boldsymbol{y}_0)$ in (20) to obtain the optimal values $t_\star, \boldsymbol{p}_\star, \boldsymbol{q}_\star, \boldsymbol{x}_\star, \boldsymbol{y}_\star$
   2.3. Compute $\texttt{error} = |t_\star - t_{\text{old}}|$
   2.4. Update $t_{\text{old}} = t_\star, \boldsymbol{x}_0 = \boldsymbol{x}_\star, \boldsymbol{y}_0 = \boldsymbol{y}_\star$.

Then the problem (19) can be approximated as

$$\boldsymbol{Q}_2(\boldsymbol{x}_0, \boldsymbol{y}_0): \quad \underset{t,\boldsymbol{p},\boldsymbol{q},\{x_k,y_k\}_{k=1}^{K_C}}{\texttt{minimize}} \quad t \tag{20}$$
$$\texttt{s.t.} \quad (18a), (18c), (19a), (19d), (19e),$$

where $\boldsymbol{x}_0 \triangleq \{x_{0k}\}_{k=1}^{K_C}, \boldsymbol{y}_0 \triangleq \{y_{0k}\}_{k=1}^{K_C}$.

For a known feasible set $\{x_{0k}, y_{0k}\}_{k=1}^{K_C}$, it is evident that problem (20) is convex, since the objective function and the constraints are convex. Hence, standard methods can be used to solve this problem effectively. We note that the approximated problem (20) gives a suboptimal solution of problem (19) because $e^{x_{0k}}(x_k - x_{0k} + 1) \leq e^{x_k}, \forall x_{0k}$.

Since the optimal solution of problem (20) is influenced by the parameters $\boldsymbol{x}_0, \boldsymbol{y}_0$. An iterative optimization algorithm is proposed in Tab. III to improve the performance of the approximated problem (20). The convergence of the proposed iterative algorithm is given in the following proposition.

Proposition 2: The objective function of problem $\boldsymbol{Q}_2(\boldsymbol{x}_0, \boldsymbol{y}_0)$ in (20) solved by the iterative algorithm in Table III decreases by iterations.

Proof: See Appendix B. ∎

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.

B. Delivery time minimization under MMSE design

The precoding vector under the MMSE design is given as $\boldsymbol{v}_k = \sqrt{p_k}\tilde{\mathbf{h}}_k$, where $\tilde{\mathbf{h}}_k$ is the k-th column of the MMSE precoding matrix $\boldsymbol{H}^H(\boldsymbol{H}\boldsymbol{H}^H + \sigma^2\boldsymbol{I})^{-1}$. Substituting $\boldsymbol{v}_k$ into (3) and (5), we obtain:

$$C_{\texttt{coop},k}^{MSE} = W\log_2\Big(1+\frac{|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_k|^2 q_k}{\sum\limits_{k\neq l\in\mathcal{U}_C}|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_l|^2 q_l + \eta\sum\limits_{i=1}^{K}\|\tilde{\boldsymbol{h}}_i\|^2 p_i + \sigma^2}\Big), \forall k \in \mathcal{U}_C$$
$$R_{\texttt{coop},k}^{MSE} = W\log_2\Big(1+\frac{|\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_k|^2 p_k}{\sum\limits_{i\neq k}|\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_i|^2 p_i + \sigma^2}\Big), \forall k.$$

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

$$\underset{\{p_k\}_{k=1}^{K},\{q_k\}_{k=1}^{K_C}}{\texttt{minimize}} \quad \max\Big(\frac{Q}{R_{\texttt{coop},1}^{MSE}}, \ldots, \frac{Q}{R_{\texttt{coop},K}^{MSE}}\Big), \tag{21}$$
$$\texttt{s.t.} \quad C_{\texttt{coop},k}^{MSE} \geq \bar{\mu}_k R_{\texttt{coop},k}^{MSE}, \forall k \in \mathcal{U}_C \tag{21a}$$
$$\sum\nolimits_{k\in\mathcal{U}_C} \|\tilde{\boldsymbol{w}}_k\|^2 q_k \leq P_{BS} \tag{21b}$$
$$\sum\nolimits_{k=1}^{K} \|\tilde{\boldsymbol{h}}_k\|^2 p_k \leq KP_{EN}. \tag{21c}$$

By using $C_{\texttt{coop},k}^{MSE}, R_{\texttt{coop},k}^{MSE}$ and introducing a new variable $t = \max\Big(\frac{Q}{R_{\texttt{coop},1}^{MSE}}, \ldots, \frac{Q}{R_{\texttt{coop},K}^{MSE}}\Big)$, we can reformulated problem (21) as follows:

$$\underset{t,\{p_k,q_l\}}{\texttt{minimize}} \quad t \tag{22}$$
$$\texttt{s.t.} \quad \log\Big(1+\frac{|\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_k|^2 p_k}{\sum_{i\neq k}|\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_i|^2 p_i+\sigma^2}\Big) \geq \frac{Q\log(2)}{tW}, \forall k \tag{22a}$$
$$\log\Big(1+\frac{|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_k|^2 q_k}{\sum\limits_{l\neq k}|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_l|^2 q_l+\eta\sum\limits_{i=1}^{K}\|\tilde{\boldsymbol{h}}_i\|^2 p_i+\sigma^2}\Big) \geq$$
$$\bar{\mu}_k\log\Big(1+\frac{|\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_k|^2 p_k}{\sum_{i\neq k}|\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_i|^2 p_i+\sigma^2}\Big), \forall k \in \mathcal{U}_C \tag{22b}$$
$$(21b), (21c).$$

In the next step, lets define parameters $A_{1k}, A_{2k}, \boldsymbol{\lambda}$ as in Sec. V-A, and following parameters:

$$E_{1k} = [|\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_1|^2, \ldots, |\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_K|^2, \sigma^2]$$
$$E_{2k} = [|\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_1|^2, \ldots, |\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_{k-1}|^2, 0, |\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_{k+1}|^2, \ldots, |\boldsymbol{h}_k^H\tilde{\boldsymbol{h}}_K|^2, \sigma^2]$$
$$\boldsymbol{\beta} = [\|\tilde{\boldsymbol{h}}_1\|^2, \ldots, \|\tilde{\boldsymbol{h}}_K\|^2, 0].$$

Then, the problem 22 can be reformulated as follows:

$$\underset{t,\boldsymbol{q},\boldsymbol{p}}{\texttt{minimize}} \quad t \tag{23}$$
$$\texttt{s.t.} \quad \log(E_{1k}\boldsymbol{p}) \geq \frac{Q\log(2)}{Wt} + \log(E_{2k}\boldsymbol{p}), \forall k \tag{23a}$$
$$\log(A_{1k}\boldsymbol{q} + \eta\boldsymbol{\beta}\boldsymbol{p}) + \bar{\mu}_k\log(E_{2k}\boldsymbol{p})$$
$$\geq \bar{\mu}_k\log(E_{1k}\boldsymbol{p}) + \log(A_{2k}\boldsymbol{q} + \eta\boldsymbol{\beta}\boldsymbol{p}), \forall k \in \mathcal{U}_C \tag{23b}$$
$$\boldsymbol{\lambda}\boldsymbol{q} \leq P_{BS};\ \boldsymbol{\beta}\boldsymbol{p} \leq KP_{EN}, \tag{23c}$$

where $\boldsymbol{p} = [p_1, \ldots, p_K, 1]^T$ and $\boldsymbol{q} = [q_1, \ldots, q_{K_C}, 1]^T$.

We observe that problem (23) is in a similar form as problem (13), except the last constraint on the EN's transmit power. Since this constraint is linear, hence convex, we can employ the same technique in Sec. IV to solve (23). Obviously, the convergence of the iterative optimization algorithm solving (23) is guaranteed by Proposition 1.

C. Delivery time minimization under optimal precoding design

In this subsection, we minimize the delivery time via general (and optimal) precoding design on the access links which jointly optimizes both direction and magnitude of the beamforming vectors $\boldsymbol{v}_k \in \mathbb{C}^{K\times 1}, \forall k$. The backhaul and access rate in this case are given as

$$C_{\texttt{coop},k}^{Opt} = W\log_2\Big(1+\frac{|\boldsymbol{g}^H\tilde{\boldsymbol{w}}_k|^2 q_k}{\sum\limits_{k\neq l\in\mathcal{U}_C}|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_l|^2 q_l+\eta\sum\limits_{i=1}^{K}\|\boldsymbol{v}_k\|^2+\sigma^2}\Big), \forall k \in \mathcal{U}_C$$
$$R_{\texttt{coop},k}^{Opt} = W\log_2\Big(1+\frac{|\boldsymbol{h}^H\boldsymbol{v}_k|^2}{\sum\limits_{i\neq k}|\boldsymbol{h}_k^H\boldsymbol{v}_i|^2+\sigma^2}\Big), \forall k.$$

The delivery time minimization problem under the optimal design is formulated as follows:

$$\underset{\{\boldsymbol{v}_k\}_{k=1}^{K},\{q_k\}_{k=1}^{K_C}}{\texttt{minimize}} \quad \max\left(\frac{Q}{R^{Opt}_{\texttt{coop},1}},\ldots,\frac{Q}{R^{Opt}_{\texttt{coop},K}}\right), \tag{24}$$

$$\texttt{s.t.} \quad C^{Opt}_{\texttt{coop},k} \geq \bar{\mu}_k R^{Opt}_{\texttt{coop},k}, \forall k \in \mathcal{U}_C \tag{24a}$$

$$\sum\nolimits_{k\in\mathcal{U}_C} \| \tilde{\boldsymbol{w}}_k \|^2 q_k \leq P_{BS} \tag{24b}$$

$$\sum\nolimits_{k=1}^{K} \|\boldsymbol{v}_k\|^2 \leq K P_{EN}. \tag{24c}$$

The challenge in solving (24) lies in the appearance of $\|\boldsymbol{v}_k\|^2$ in the denominator of both backhaul and access rates. To leverage this difficulty, we introduce new variables $\boldsymbol{V}_k \triangleq \boldsymbol{v}_k^H \boldsymbol{v}_k \in \mathbb{C}^{K\times K}$, which is symmetric and positive definite. It is straightforward to verify that $\|\boldsymbol{v}_k\|^2 = \mathrm{Tr}(\boldsymbol{V}_k)$ and $|\boldsymbol{h}_k^H \boldsymbol{v}_i|^2 = \mathrm{Tr}(\boldsymbol{H}_k \boldsymbol{V}_i)$, where $\boldsymbol{H}_k \triangleq \boldsymbol{h}_k^H \boldsymbol{h}_k$. Furthermore, by using a slack variable t we can equivalently reformulate problem (24) similarly to the previous subsection as

$$\underset{t,\{\boldsymbol{V}_k,q_l\}}{\texttt{minimize}} \quad t \tag{25}$$

$$\texttt{s.t.} \quad \log\left(1+\frac{\mathrm{Tr}(\boldsymbol{H}_k\boldsymbol{V}_k)}{\sum_{i\neq k}\mathrm{Tr}(\boldsymbol{H}_k\boldsymbol{V}_i)+\sigma^2}\right) \geq \frac{Q\log(2)}{tW}, \forall k \tag{25a}$$

$$\log\left(1+\frac{|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_k|^2 q_k}{\sum\limits_{l\neq k}|\boldsymbol{g}_k^H\tilde{\boldsymbol{w}}_l|^2 q_l+\sum\limits_{i=1}^{K}\mathrm{Tr}(\boldsymbol{V}_i)+\sigma^2}\right) \geq$$

$$\bar{\mu}_k \log\left(1+\frac{\mathrm{Tr}(\boldsymbol{H}_k\boldsymbol{V}_k)}{\sum_{i\neq k}\mathrm{Tr}(\boldsymbol{H}_k\boldsymbol{V}_i)+\sigma^2}\right), \forall k \in \mathcal{U}_C \tag{25b}$$

$$\sum_{k=1}^{K}\mathrm{Tr}(\boldsymbol{V}_k) \leq K P_{EN};\ \mathrm{rank}(\boldsymbol{V}_k)=1, \forall k, \tag{25c}$$

(24b).

By using similar notations $A_{1k}, A_{2k}, \boldsymbol{\lambda}$ as in Sec. V-A, we can reformulate (25) as

$$\underset{t,\{\boldsymbol{V}_k\},\boldsymbol{q}}{\texttt{minimize}} \quad t \tag{26}$$

$$\texttt{s.t.} \quad \log(\textstyle\sum_{i=1}^{K}\mathrm{Tr}(\boldsymbol{H}_k\boldsymbol{V}_i)+\sigma^2) \geq \tag{26a}$$

$$\frac{Q\log(2)}{tW}+\log(\textstyle\sum_{i\neq k}\mathrm{Tr}(\boldsymbol{H}_k\boldsymbol{V}_i)+\sigma^2), \forall k$$

$$\log(A_{1k}\boldsymbol{q}+\eta\sum_{i=1}^{K}\mathrm{Tr}(\boldsymbol{V}_i))+\bar{\mu}\log(\sum_{i\neq k}\mathrm{Tr}(\boldsymbol{H}_k\boldsymbol{V}_i)+\sigma^2) \geq$$

$$\log(A_{2k}\boldsymbol{q}+\eta\sum_{i=1}^{K}\mathrm{Tr}(\boldsymbol{V}_i))+\bar{\mu}\log(\sum_{i=1}^{K}\mathrm{Tr}(\boldsymbol{H}_k\boldsymbol{V}_i)+\sigma^2) \tag{26b}$$

$$\boldsymbol{\lambda}\boldsymbol{q} \leq P_{BS};\ \sum_{k}\mathrm{Tr}(\boldsymbol{V}_k) \leq K P_{EN} \tag{26c}$$

$$\mathrm{rank}(\boldsymbol{V}_k)=1, \forall k,$$

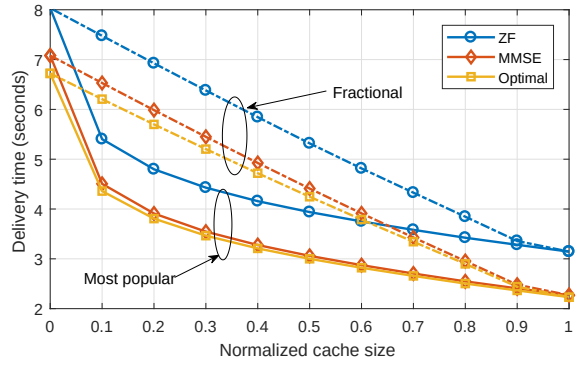
where constraint (26b) is applied only for $k \in \mathcal{U}_C$.

Solving problem (26) is difficult due to the non-convexity of (26a), (26b) and the rank-one constraint. In order to deal with the latter, we employ the semidefinite relaxation (SDR) method [31] which ignores the rank-one constraint when solving (26). SDR has been widely known as an efficient solution that achieves a close performance to the optimum [31][2]. To over the former, we observe that the trace function is linear and (26a) and (26b) are in similar form as constraints (13a) and (13b), respectively. Therefore, we can can employ similar technique in Section IV to solve the SDR of (26), whose details are skipped to avoid redundancy.

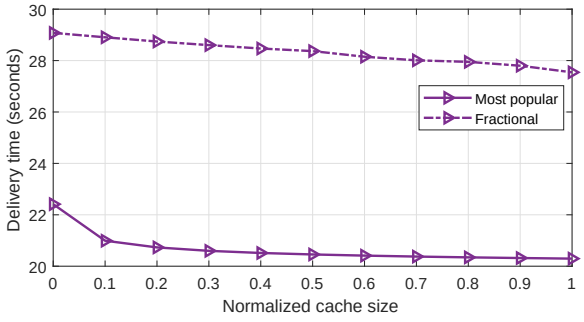
[2]Since the SDR solution does not always guarantee the rank-one constraint, Gaussian randomization can be applied to improve the final performance. Details of Gaussian randomization technique are available in [31].

VI. Numerical results

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 is $G_1 = -60$dB. The pathloss on the access intended links is $G_2 = -50$dB. The pathloss on the inter-EN channels, e.g., f_{kl}, and the access interfering links, e.g., $h_{ki}, i \neq k$, are $G_E = -56$dB. Unless stated otherwise, the self-interference cancellation efficiency is equal to $\bar{\eta} = -70$dB [25]. Other parameters are as follows: $N = K = 4$, $\sigma^2 = -100$ dBm, $F = 100$ files, $Q = 100$Mb, and $W = 10$MHz. The simulation results are calculated based on 10000 random requests, equally distributed over 200 channel realizations. To achieve the best performance, we run the proposed iterative algorithms with 100 different initial values (see Table II and III for details) and select the best value. The user requests are assumed to follow the Zipf distribution with the skewness factor $\xi = 0.8$. In the figures, we use ZF, MMSE and Optimal to refer to ZF, MMSE and Optimal precoding designs, respectively.

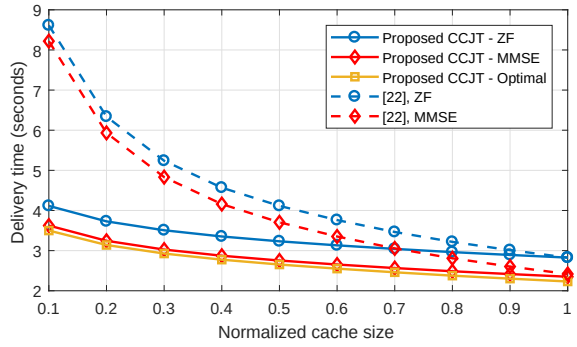


(a) CCJT scheme

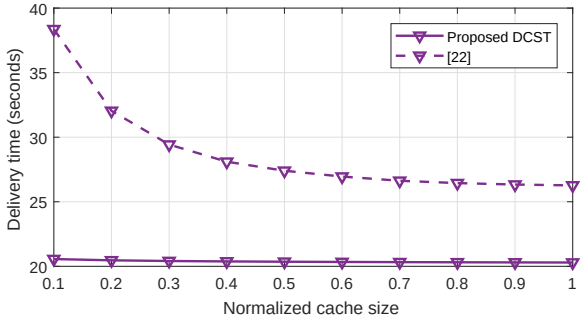


(b) DCST scheme

Fig. 2: Performance comparison of the proposed schemes under both most popular and fractional caching policies. $P_{BS} = 3.16$W (5 dB) and $P_{EN} = 5$W.



(a) CCJT scheme



(b) DCST scheme

Fig. 3: The caching gain in FD-MEC systems v.s. the WAP's transmit power. EN's transmit power $P_{EN} = 5$W, the WAP's transmit pwoer $P_{BS} = 10$W.

A. Most popular caching versus fractional caching

Fig. 2 presents the delivery time performance of the proposed CCJT (a) and DCST (b) as a function of the normalized cache size, the ratio of the cache size divided by the library size, i.e., $\frac{M}{F}$. Both the most popular and fractional caching policies are presented. In the former, the most M popular files are prefetched in the EN's cache, while in the later, a portion $\frac{M}{F}$ of every files are cached. In general, the most popular caching policy spends less time to serve the user requests than the fractional caching in both CCJT and DCST schemes. This is because the user requests follow a Zipf-based distribution, in which popular files are requested more frequently than the less popular ones. Since the most popular caching policy is more efficient than the fractional caching strategy, we only present the results for the most popular caching in the rest of the paper.

B. Effectiveness of the proposed optimization algorithms and cooperative caching

We compare the proposed FD-MEC optimization algorithms with [22], which proposes a FD-aided edge caching scheme with static transmit power. Although [22] considers only DCST, this method can be directly applied to CCJT under linear precoding designs without power control. In Fig. 3, we demonstrate the effectiveness of the proposed optimization algorithms in both CCJT (a) and DCST (b). It is noted that the reference [22] under linear precoding designs, i.e., ZF and MMSE, always transmit at the maximum power, equally divided for the ENs on the backhaul and for users on the access channels. A large gain is observed for the proposed optimization algorithms compared to the reference, especially in the small and medium cache size regimes. At large cache sizes, most of the requested files will be available in the ENs' cache, hence less traffic on the backhaul is required. In this case, the equal-power mode achieves a close performance as the proposed scheme. Consider the precoding designs in CCJT, the MMSE design performs considerably better than the ZF and achieves a close performance to the optimal precoding design. This is because MMSE and Optimal schemes perform power allocation more effectively than ZF, especially when the channel matrix is low rank. On average, the ZF design spends one second more than the tow others to serve the same demands. From a practical perspective, MMSE is preferred due to its low computation complexity compared with the Optimal scheme, as shown in Table IV.

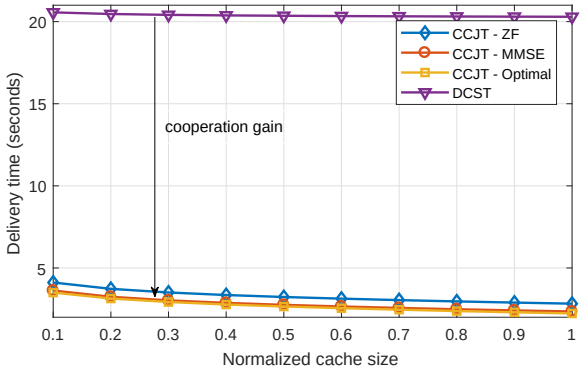


Fig. 4: The caching gain in FD-MEC systems v.s. the WAP's transmit power. EN's transmit power $P_{EN} = 5$W, cache size $M = 0.4F$.

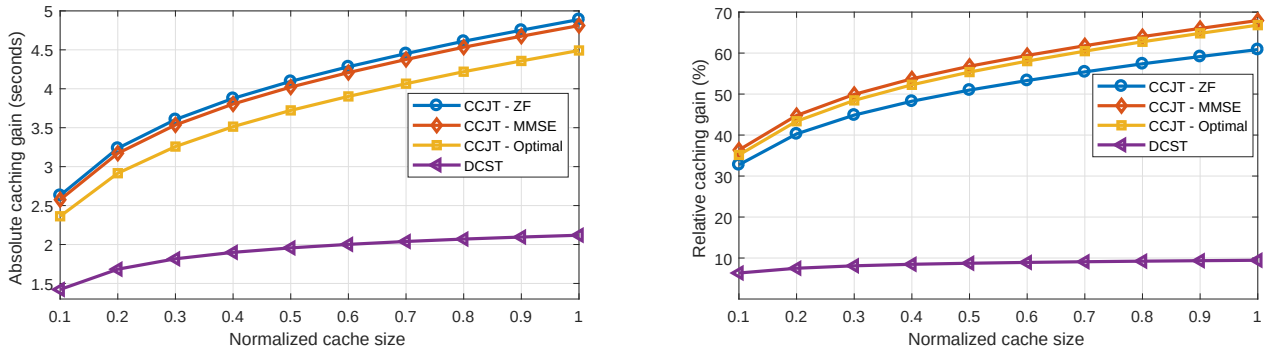
TABLE IV: Average simulation time (in seconds) of three precoding designs, $K = 4$.

ZF	MMSE	Optimal
0.0409	0.0509	0.1499

Fig. 4 compares the delivery time of the CCJT with the DCST modes as a function of the normalized cache size. We recall that the DCST is fully decentralized and each EN operates independently. By allowing cooperative caching and joint transmission among the ENs, the delivery time dramatically drops for all cache sizes. In particular, the CCJT reduces the delivery time by about 85% compared with DCST, which is mainly limited by both inter-EN and self interference. Obviously, this gain comes at the expense of extra physical connection and signal overheads among the ENs.

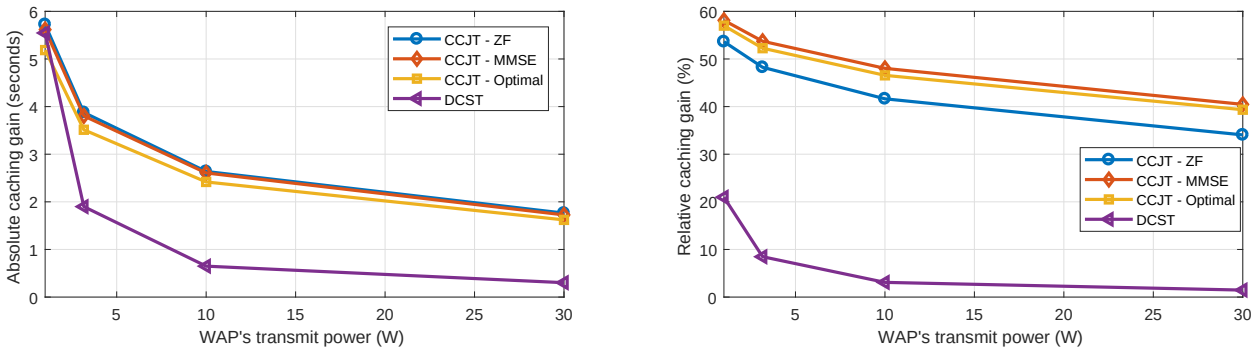
C. Role of caching in FD-MEC systems

The effectiveness of caching in FD systems is demonstrated via a caching gain metric, which is computed as the delivery time reduction brought by the FD-MEC compared with the FD systems without caching capability at the ENs. In order to



(a) Absolute caching gain (b) Relative gain

Fig. 5: The caching gain in FD-MEC systems v.s. the normalized cache size $\frac{M}{F}$. $P_{BS} = 3.16$W (5 dB) and $P_{EN} = 5$W. Solid lines show the most popular caching policy. Dotted lines show the fractional caching policy.



(a) Absolute caching gain (b) Relative gain

Fig. 6: The caching gain in FD-MEC systems v.s. the WAP's transmit power. EN's transmit power $P_{EN} = 5$W, cache size $M = 0.4F$.

provide a complete observation, two types of caching gain are presented: *Absolute caching gain* (ACG) and *Relative caching gain* (RCG), which is calculated as follows:

$$\text{ACG} = t_{\text{no cache}} - t_{\text{cache}};\ \text{RCG} = 1 - \frac{t_{\text{cache}}}{t_{\text{no cache}}},$$

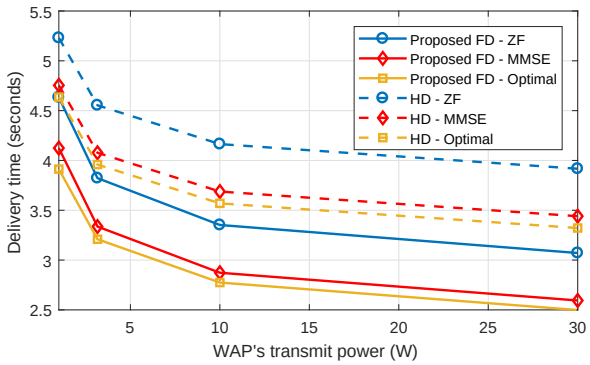
where t_{cache} and $t_{\text{no cache}}$ are the delivery time of the FD systems with and without caching at the ENs, respectively.

Fig. 5 presents the absolute caching gain (a) and relative caching gain (b) of the FD-MEC versus the normalized cache size. In general, caching in the cooperative mode CCJT is significantly more efficient than in the distributed architecture DCST. This expected outcome originates from two reasons. First, the shared cache among the ENs in CCJT facilitates the self-interference cancellation on the backhaul. Second, the joint transmission on the access links undoubtedly improves the access rates. At the cache size $M = 0.5F$, the CCJT (with all designs) achieves about 4 seconds reduction of the delivery time, twice as the DCST (Fig. 5a). The role of caching is shown more clearly via the relative caching gain in Fig. 5b: it reduces the delivery time by 55% in the CCJT, compared with only 10% in the DCST. We note that having the normalized cache size equal 1, i.e., $M = F$, does not result in 100% relative caching gain since the total delivery time is lower bounded by the access channels. It is noted that although the ZF design achieves a larger absolute caching gain than MMSE and the Optimal, its relative gain is smaller. This implies that the ZF design is less efficient than the others.

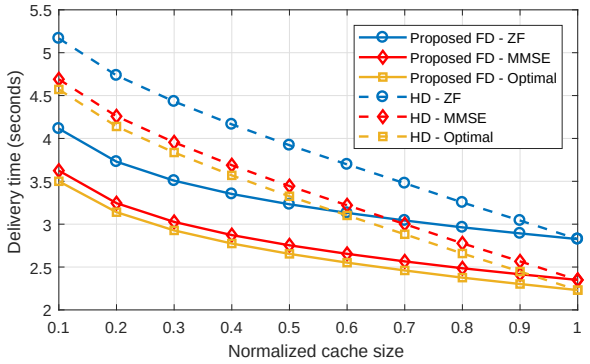
Fig. 6 shows the caching gains as a function of the WAP's transmit power, with $M = 0.4F$ and $P_{EN} = 5$W. A similar conclusion is drawn that CCJT is much more efficient than DCST. In addition, the influence of WAP's transmit power on the caching gain reduces as P_{BS} becomes large. This is because at large WAP transmit powers, the delivery time is mainly determined by the access channel quality.

D. Comparison with half-duplex systems

In HD systems, the backhaul and access transmissions 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. 7a plots the delivery time as a function of the WAP's transmit power P_{BS} under the CCJT mode, with $M = 0.4F$ and $P_{EN} = 5$W. It is observed that the FD-MEC system largely reduces the delivery time compared with the HD scheme for all precoding designs, i.e., ZF, MMSE and Optimal. At the WAP's transmit power equal to 5W, a reduction of 25% is obtained by the FD scheme for all



(a) v.s. WPA's transmit power. $M = 0.4F$



(b) v.s. normalized cache size. $P_{BS} = 10$W

Fig. 7: Delivery time comparison between FD-MEC and HD system under the CCJT architecture. $P_{EN} = 5$W.

precoding designs. An important observation is that large values of P_{BS} will have less influence 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 given a finite P_{EN}.

Fig. 7b compares the delivery time of the FD-MEC with the HD system under the CCJT mode versus the normalized cache size, i.e., $\frac{M}{F}$. It is shown that the gain offered by the FD system over the HD is more significant in the small cache size ranges. The benefit of caching can be also interpreted as a means of trading memory for power: the delivery time with a large transmit power ($P_{BS} = 30$W, $M = 0.4F$ in Fig. 7a) can also be achieved with a smaller transmit power and a larger cache size ($P_{BS} = 10$W, $M = 0.6F$ in Fig. 7b). Increasing the cache size will diminish the advantage of the FD scheme over the HD. As such, 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. 8 plots the delivery time versus the self-interference cancellation efficiency $\bar{\eta}$. 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 is shown that the FD system outperforms the HD mode in the small values of $\bar{\eta}$. When the performance of the interference

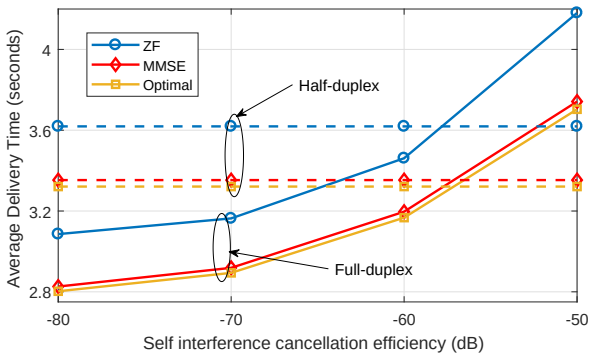


Fig. 8: Average delivery time v.s. the self-interference cancellation efficiency $\bar{\eta}$. $M = 0.4F$, $P_{BS} = 10$W, and $P_{EN} = 5$W.

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 a cache-aided system.

VII. Conclusions

In this paper, we have investigated the performance of full-duplex enabled mobile edge caching systems via the delivery time metric. The considered system is analysed under two network architectures: distributed caching and cooperative caching. For each architecture, we proposed an optimal power control to minimize the system delivery time based on the linear precoding design. To overcome the non-convexity of the formulated problems, two iterative optimization algorithms have been proposed based on the inner approximation method, whose convergence is analytically guaranteed. We have demonstrated that the cooperative caching perform largely better than the distributed scheme at the expense of full cooperation among the ENs. It has been also shown that the MMSE-based precoding design achieves the best trade-off between the performance and computation complexity.

The considered schemes represent the two extremes of FD-MEC systems when collaboration among the ENs is available: i) the ENs operate in a complete decentralized manner, and ii) the ENs fully cooperate. Practical scenarios usually fall between these two modes. In this case, a cluster of ENs collaborate to serve their users, while the rest of the ENs operate independently. One promising extension from this work is to optimize the caching policy at the ENs. This would require the derivation of the average delivery time over all fading channels.

Appendix A
Proof of Proposition 1

Denote $\left(t_\star^{(i)}, \boldsymbol{p}_\star^{(i)}, \boldsymbol{q}_\star^{(i)}, \boldsymbol{x}_\star^{(i)}, \boldsymbol{y}_\star^{(i)}, \boldsymbol{z}_\star^{(i)}\right)$ as the optimal solution of $\boldsymbol{Q}_1(\boldsymbol{x}_0^{(i)}, \boldsymbol{y}_0^{(i)}, \boldsymbol{z}_0^{(i)})$ at iteration i. We will show that if $x_{\star k}^{(i)} < x_{0k}^{(t)}, \forall k$, then by using $x_{0k}^{(i+1)} = x_{\star k}^{(i)}$ in the $(i+1)$-th iteration, we will have $t_\star^{(i+1)} < t_\star^{(i)}$. Indeed, by choosing a relatively large initial value $\boldsymbol{x}_0^{(1)}, \boldsymbol{y}_0^{(1)}, \boldsymbol{z}_0^{(1)}$, we always have $x_{\star k}^{(1)} < x_{0k}^{(1)}, \forall k$.

Denote $f(x;a) = e^a(x-a+1)$ as the first order approximation of function e^x 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, $f(x; x_{\star k}^{(i)})$ is used in the right-hand side of constraint (15a). Consider a candidate $\boldsymbol{x}^{(i+1)} = \{x_1^{(i+1)}, \ldots, x_{K_C}^{(i+1)}\}$, with $x_k^{(i+1)} = x_{\star k}^{(i)} - 1 + e^{x_{0k}^{(i)} - x_{\star k}^{(i)}}(x_{\star k}^{(i)} - x_{0k}^{(i)} + 1)$. It is evident that $x_k^{(i+1)} < x_{\star k}^{(i)}$ and $f(x_k^{(i+1)}; x_{\star k}^{(i)}) = f(x_{\star k}^{(i)}; x_{0k}^{(i)}), \forall k \leq K_C$.

Because $x_k^{(i+1)} < x_{\star k}^{(i)}, \forall k \leq K_C$, the strictly inequality holds in constraint (14a). Thus, there exits $t^{(i+1)} < t_\star^{(i)}$ which satisfies $\log(A_k \boldsymbol{p}) \geq \frac{Q\log(2)}{t^{(i+1)}} + x_k^{(i+1)}, \forall k$. Now consider a new candidate set $(t^{(i+1)}, \boldsymbol{p}_\star^{(i)}, \boldsymbol{q}_\star^{(i)}, \boldsymbol{x}^{(i+1)}, \boldsymbol{y}_\star^{(i)}, \boldsymbol{z}_\star^{(i)})$. This set satisfies all the constraints of problem $\boldsymbol{Q}_1(\boldsymbol{x}_\star^{(i)}, \boldsymbol{y}_0^{(i)}, \boldsymbol{z}_0^{(i)})$, and therefore is a feasible solution of the optimization problem. As a 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.

Appendix B
Proof of Proposition 2

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 $\boldsymbol{Q}_2(\boldsymbol{x}_0^{(i)}, y_0^{(i)})$ at iteration i. We will show that if $x_{\star k}^{(i)} < x_{0k}^{(i)}$ and $y_\star^{(i)} > y_0^{(i)}, \forall k \leq K_C$, then by using $x_{0k}^{(i+1)} = x_{\star k}^{(i)}$, $y_{0k}^{(i+1)} = y_{\star k}^{(i)}$ in the $(i+1)$-th iteration, we will have $t_\star^{(i+1)} < t_\star^{(i)}$. Indeed, by choosing a relatively large initial value $\{x_0^{(1)}\}_{k=1}^{K_C}$ and small value $\{y_{0k}^{(1)}\}_{k=1}^{K_C}$, we always have $x_{\star k}^{(1)} < x_{0k}^{(1)}$ and $y_{\star k}^{(1)} > y_{0k}^{(1)}, \forall k \leq K_C$.

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, $f(x; x_{\star k}^{(i)})$ is used in the right-hand side of constraint (19d), where $f(x;a) = e^a(x-a+1)$ is the first order approximation at a of function e^x. 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 $f(x_k^{(i+1)}; x_{\star k}^{(i)}) > f(x_{\star k}^{(i)}; x_{0k}^{(i)}), \forall k \leq K_C$. In addition, consider a candidate $y^{(i+1)} = y_\star^{(i)} + \delta y$, with $\delta y \leq \min_{1 \leq k \leq K_C}\{\bar{\mu}_k(x_{\star k}^{(i)} - x_k^{(i+1)})\}$. Obviously, $f(y_k^{(i+1)}; y_{\star k}^{(i)}) > f(y_{\star k}^{(i)}; y_{0k}^{(i)})$ due to the convexity of e^y function.

Because $f(x_k^{(i+1)}; x_{\star k}^{(i)}) > f(x_{\star k}^{(i)}; x_{0k}^{(i)})$ and $f(y_k^{(i+1)}; y_{\star k}^{(i)}) > f(y_{\star k}^{(i)}; y_{0k}^{(i)}), \forall k \leq K_C$, the strict inequality holds in constraints (19d) and (19e). Thus, there exits $p_k^{(i+1)} > p_{\star k}^{(i)}$ and $t^{(i+1)} < t_\star^{(i)}$ which satisfies constraints (18a), (19d) and (19e). Furthermore, since $\delta y \leq \min_{1 \leq k \leq K_C}\{\bar{\mu}_k(x_{\star k}^{(i)} - x_k^{(i+1)})\}$, constraint (19a) is also satisfied. Now consider a new candidate set $(t^{(i+1)}, \boldsymbol{p}^{(i+1)}, \boldsymbol{q}_\star^{(i)}, \boldsymbol{x}^{(i+1)}, \boldsymbol{y}^{(i+1)})$. This set satisfies all the constraints of problem $\boldsymbol{Q}_2(\boldsymbol{x}_\star^{(i)}, \boldsymbol{y}_\star^{(i)})$, and therefore is a feasible solution of the optimization problem. As a 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 2.

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 optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2827 – 2839, Apr. 2018.

[4] T. X. Vu, S. Chatzinotas, B. Ottersten, and T. Q. Duong, "Energy minimization for cache-assisted content delivery networks with wireless backhaul," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 332 - 335, Jun. 2018.

[5] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sept. 2016.

[6] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[7] T. X. Tran and D. Pompili, "Octopus: A cooperative hierarchical caching strategy for cloud radio access networks," in *Proc. Int. Conf. Mobile Ad-Hoc Sensor Syst.*, Brasilia, Oct. 2016, pp. 154–162.

[8] 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.

[9] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. Annu. Conf. Info. Sci. Syst.*, Princeton, NJ, Mar. 2016, pp. 320–325.

[10] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438–2452, Jun. 2016.

[11] M. Gregori, J. Gmez-Vilardeb, J. Matamoros, and D. Gndz, "Wireless content caching for small cell and D2D networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[12] 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.

[13] 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.

[14] 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.

[15] L. Lei, E. Lagunas, S. Chatzinotas, and B. Ottersten, "NOMA aided interference management for full-duplex self-backhauling hetnets," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1696–1699, Aug. 2018.

[16] L. T. Tan, R. Q. Hu, and Y. Qian, "D2D communications in heterogeneous networks with full-duplex relays and edge caching," *IEEE Trans. Industrial Informatics (Special issue on Fog Computing for Industrial Applications)*, vol. 14, no. 10, pp. 4557–4567, Oct. 2018.

[17] S. K. Sharma et al, "Dynamic spectrum sharing in 5G wireless networks with full-duplex technology: Recent advances and research challenges," *IEEE Commun. Surveys & Tutorials*, vol. 20, no. 1, PP. 674 – 707, 2018.

[18] L. Zhang, M. Xiao, G. Wu, M. Alam, YC Liang, and S. Li, "A survey of advanced techniques for spectrum sharing in 5G networks," *IEEE Wireless Commun. Mag.,* vol. 24, no. 5, pp. 44–51, 2017.

[19] 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.

[20] 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.

[21] T. X. Vu, L. Lei, S. Chatzinotas, B. Ottersten and A. V. Trinh, "On the Successful Delivery Probability of Full-Duplex-Enabled Mobile Edge Caching," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 1016–1020, Jun. 2019.

[22] M. Maso, I. Atzeni, I. Ghamnia, E. Batu, and M. Debbah, "Cache-aided full-duplex small cells," in *15th Int. Symp. on Modeling and Opt. in Mobile, Ad Hoc, and Wireless Netw. (WiOpt)*, May 2017, pp. 1–6.

[23] 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.

[24] M. E. Knox, "Single antenna full duplex communications using a common carrier," in *WAMICON 2012 IEEE Wireless Microwave Technology Conference*, Apr. 2012, pp. 1–6.

[25] D. Bharadia and S. Katti, "Full duplex MIMO radios," in *Proc. 11th USENIX Conf. Netw. Sys. Design and Implementation*, ser. NSDI'14, no. 14. Berkeley, CA, USA: USENIX Association, 2014, pp. 359–372.

[26] 3GPP TS 36.423, "Evolved universal terrestrial radio access network (e-utran); x2 application protocol (x2ap)," release 8.
[27] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *IEEE INFOCOM*, vol. 1, March 1999, pp. 126–134 vol.1.
[28] D. Bharadia and S. Katti, "Fastforward: Fast and constructive full duplex relays," in *Proc. 2014 ACM Conf. on SIGCOMM*. ACM, 2014, pp. 199–210.
[29] N. H. Mahmood, I. S. Ansari, G. Berardinelli, P. Mogensen, and K. A. Qaraqe, "Analysing self interference cancellation in full duplex radios," in *Proc. IEEE Wireless Commun. Netw. Conf.*, April 2016, pp. 1–6.
[30] T. X. Vu, T. A. Vu, L. Lei, S. Chatzinotas, and B. Ottersten, "Linear precoding design for cache-aided full-duplex networks," *IEEE Wireless Commun. Netw. Conf.*, 2019, pp. 1–6.
[31] Z.-Q. Luo, W. K. Ma, A. M. C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, Mar. 2010.

Thang X. Vu (S'11–M'15) was born in Hai Duong, Vietnam. He received the B.S. and the M.Sc., both in Electronics and Telecommunications Engineering, from the VNU University of Engineering and Technology, Vietnam, in 2007 and 2009, respectively, and the Ph.D. in Electrical Engineering from the University Paris-Sud, France, in 2014.

From 2007 to 2009, he was with the Department of Electronics and Telecommunications, VNU University of Engineering and Technology, Vietnam as a research assistant. In 2010, he received the Allocation de Recherche fellowship to study Ph.D. in France. From September 2010 to May 2014, he was with the Laboratory of Signals and Systems (LSS), a joint laboratory of CNRS, CentraleSupelec and University Paris-Sud XI, France. From July 2014 to January 2016, he was postdoctoral researcher with the Information Systems Technology and Design (ISTD) pillar, Singapore University of Technology and Design (SUTD), Singapore. Currently, he is research associate at Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg. His research interests are in the field of wireless communications, with particular interests of cache-assisted 5G, cloud radio access networks, resources allocation and optimization, cooperative diversity, channel and network decoding, and iterative decoding.

Symeon Chatzinotas (S'06-M'09-SM'13) is currently the Deputy Head of the SIGCOM Research Group, Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg, Luxembourg and Visiting Professor at the University of Parma, Italy. He received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, Surrey, U.K., in 2006 and 2009, respectively. He was involved in numerous Research and Development projects for the Institute of Informatics Telecommunications, National Center for Scientific Research Demokritos, the Institute of Telematics and Informatics, Center of Research and Technology Hellas, and the Mobile Communications Research Group, Center of Communication Systems Research, University of Surrey. He has over 250 publications, 2000 citations, and an H-Index of 25 according to Google Scholar. His research interests include multiuser information theory, co-operative/cognitive communications, and wireless networks optimization. He was a co-recipient of the 2014 Distinguished Contributions to Satellite Communications Award, and the Satellite and Space Communications Technical Committee, the IEEE Communications Society, and the CROWNCOM 2015 Best Paper Award.

Bjorn Ottersten (S'87-M'89-SM'99-F'04) was born in Stockholm, Sweden, in 1961. He received the M.S. degree in electrical engineering and applied physics from Linkping University, Linkping, Sweden, in 1986, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1990. He has held research positions with the Department of Electrical Engineering, Linkping University, the Information Systems Laboratory, Stanford University, the Katholieke Universiteit Leuven, Leuven, Belgium, and the University of Luxembourg, Luxembourg. From 1996 to 1997, he was the Director of Research with ArrayComm, Inc., a start-up in San Jose, CA, USA, based on his patented technology. In 1991, he was appointed a Professor of signal processing with the Royal Institute of Technology (KTH), Stockholm, Sweden. From 1992 to 2004, he was the Head of the Department for Signals, Sensors, and Systems, KTH, and from 2004 to 2008, he was the Dean of the School of Electrical Engineering, KTH. He is currently the Director for the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg.

He was a recipient of the IEEE Signal Processing Society Technical Achievement Award in 2011 and the European Research Council advanced research grant twice, in 20092013 and in 20172022. He has co-authored journal papers that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, 2006, and 2013, and seven IEEE conference papers best paper awards. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the Editorial Board of the *IEEE Signal Processing Magazine*. He is currently a member of the editorial boards of *EURASIP Signal Processing Journal, EURASIP Journal of Advances Signal Processing and Foundations and Trends of Signal Processing*. He is a fellow of EURASIP.

Anh Vu Trinh received the B.S. degree in radio physics from Hanoi University, Hanoi, Vietnam, in 1983 and the Ph.D degree in mathematics and physics from Moscow State University, Moscow, Russia, in 1994. He was a visiting researcher at Tasmania University, Australia, in 2002.

Currently, he is an Associate Professor in the Department of wireless Communications at Faculty of Electronics and Telecommunication (FET), University of Engineering and Technology (UET), Vietnam National University, Hanoi (VNU Hanoi). He is a member of the Radio Electronic Vietnam Association.

His research interests include Wireless Communications, Multi-Antenna Systems, Cahing and Computing Systems, and Implementing algorithms in FPGA.