Wireless Networks
Accepted in April 2020.

Cache-aided Full-Duplex: Delivery Time Analysis and Optimization

Thang X. Vu · Anh Vu Trinh · Symeon Chatzinotas · Xuan Nam Tran

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 improve the spectral efficiency by allowing a node to receive and transmit on the same frequency band simultaneously. In this paper, we investigate the delivery time performance of a cache-aided FD system, in which an edge node, operates in FD mode, serves users via wireless channels and is equipped with a cache memory. Firstly, we derive a closed-form expression for the average delivery time by taking into account the uncertainties of both backhaul and access wireless channels. The derived analysis allows the examination of the impact of the key parameters, e.g., cache size and transmit power. Secondly, a power optimization problem is formulated to minimize the average delivery time. To deal with the non-convexity of the formulated problem, we propose an iterative optimization algorithm based on the bisection method. Finally, numerical results are presented to demonstrate the effectiveness of the pro-

T. X. Vu, S. Chatzinotas
University of Luxembourg, Luxembourg
E-mail: {thang.vu, symeon.chatzinotas}@uni.lu

A. V. Trinh
VNU University of Engineering and Technology, Hanoi, Vietnam
E-mail: vuta@vnu.edu.vn

X. N. Tran
Le Quy Don Technical University, Hanoi, Vietnam
E-mail: namtx@mta.edu.vn

posed algorithm, which can significantly reduce the delivery time compared to the FD reference and the half-duplex counterpart.

Keywords Edge caching, delivery time, full duplex, optimization.

1 Introduction

Among potential technology enablers to tackle the stringent latency and data-hungry challenges in future wireless networks, edge caching has received much attention. By prefetching content close 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 (EN) can directly serve the users' demands without transferring the date 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 ENs when designing the signal transmission to reduce costs on both access and backhaul links. Since some (parts of) requested files are available in the EN'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 [5–8]. The performance of cache-aided wireless networks can be further improved by joint optimization of caching along with routing and resource allocation [9].

Meanwhile, full-duplex (FD) has shown great potential as the transmission technique for the next generation wireless networks [10,11]. Thanks to recent developments in the self-interference cancellation, FD can potentially double the spectral efficiency by allowing a node to transmit and receive signals at the same resource block simultaneously [12]. Despite the cache-aided HD system has been well studied in the literature, the investigation on cache-aided FD systems is limited. The authors of [13] show that cache-aided FD small cell networks can provide cache hit enhancements compared to the HD systems. Therein, by modeling the base stations and users as a coupled Poison point process (PPP) with the edge nodes, coverage probability and successful delivery rates are analysed. The role of caching in FD D2D networks is investigated in [15,16] via stochastic geometry analysis. By considering all possible operating modes of an arbitrary device, the success probability is derived in [15] as a function of the cache size and the interference distribution. It is shown in [16] that a hybrid deployment of FD and HD modes can further improve the coverage probability in cluster-based FD networks. In [17], the authors derive a closed-form expression for the successful delivery probability (SDP) of the cache-aided FD system, from which a heuristic-based caching design is proposed for SDP maximization. The worst case normalized delivery time (NDT) in heterogeneous networks is studied in [14] with the presence of FD relaying nodes. However, the result obtained in [14] is based on an optimistic assumption of perfect self-interference cancellation. In practice, there always remains residual interference after the self-interference cancellation [18,19].

1.1 Contributions

In this paper, we analyse the delivery time performance of a cache-aided FD system. We first derive a closed-form expression for the average delivery time over the wireless channel uncertainties. The derived closed-form allows us to study the contributions of the key parameters, e.g., cache size and transmit power. Next, we formulate a power optimization problem to minimize the delivery time. Since the formulated problem is non-convex, we propose an iterative optimization algorithm based on the bisection method. Numerical results verify the accuracy of our analysis and show a significant delivery time reduction compared to the FD benchmark and the half-duplex counterpart.

The rest of this paper is organized as follows. Section 2 presents the system model and the caching strategies. Section 3 analyses the average delivery time. Section 4 minimizes the system delivery time. Section 5 shows numerical results. Finally, Section 6 concludes the paper.

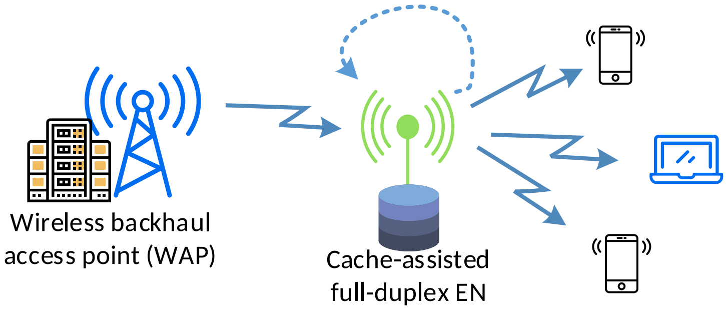
Wireless backhaul access point (WAP)
Cache-assisted full-duplex EN

Fig. 1: Cache-assisted full-duplex system. Self-interference exists at the edge node.

2 System Model

We consider a cache-aided FD system, in which an EN, operating in FD mode, serves K users via wireless channels, as depicted in Fig. 1. The EN connects to the core network via a wireless backhaul access point (WAP), e.g., macro base station. The users can only access data via the EN, i.e., there is no direct link between the users and the WAP. The WAP is assumed to have access to a library of N contents, denoted by $\mathcal{F} = \{F_1, \ldots, F_N\}$. Without loss of generality, all contents are 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 $M < N$.

2.1 Content Popularity and Caching model

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

$$q_n = C^{-1} n^{-\alpha}, \tag{1}$$

where $C = \sum_{m=1}^{N} m^{-\alpha}$ and α is the Zipf skewness factor.

We consider a generic caching policy $\mathbf{c} = [c_1, \ldots, c_N]$, where $c_n \in [0, 1]$ denotes portions of file F_n cached at the EN. In order to meet the memory constraint, it must hold that $\sum_{n=1}^{N} c_n \leq M$. The motivation behind the generic caching policy is that it allows the study of different caching strategies. Two typical use cases are Zipf-based (most popular) caching and uniform caching. In the latter, we have $\mathbf{c}_{\text{UNI}} = [\frac{M}{N}, \ldots, \frac{M}{N}]$, while in the former, we have $\mathbf{c}_{\text{Zip}} = [\underbrace{1, \ldots, 1}_{\times M}, 0, \ldots, 0]$.

2.2 Signal Transmission Model

We assume that the WAP is equipped with L antennas, while the EN and users are equipped with a single antenna [13,18]. Denote $\bar{\boldsymbol{h}} = \sqrt{G_1}\boldsymbol{h} \in \mathbb{C}^{1\times L}$ as the channel coefficients of the backhaul link, where G_1 is the path-loss degradation and $\boldsymbol{h}$ is the small-scale fading, whose elements are independent and identically distributed (i.i.d.) complex Gaussian random variables with zero-mean and unit variance. Let $\bar{g}_k = \sqrt{G_{2,k}}g_k$ denote the channel coefficient between the EN and user k, where $G_{2,k}$ is the pathloss from the EN to user k, and g_k is the fading coefficient, which is complex Gaussian random variable with zero-mean and unit variance. The K users are served by the time division multiple access (TDMA), in which each user is consecutively active in $\frac{\mathcal{T}}{K}$, where $\mathcal{T}$ is the coherence time. Therefore, there is no inter-user interference. However, since the EN operates in FD mode, there exists self-interference at the EN.

When user k requests a content, it sends the content index to the EN. The EN first checks its local cache. If (parts of) the requested content is available in the cache, it serves the user directly. Otherwise, the EN will receive the non-cached parts from the WAP via the wireless backhaul before serving the user. Let $d_k, \forall k$, denote the file index requested by user k and denote $\bar{F}_{d_k}$ as parts of file F_{d_k} which are available at the EN's cache. Denote $\breve{F}_{d_k} = F_{d_k} - \bar{F}_{d_k}$ as the non-cached parts which will be sent from the WAP. Furthermore, denote $x_{1,k}$ and $x_{2,k}$ as the modulated signal of $\breve{F}_{d_k}$ and $\bar{F}_{d_k}$, respectively. A transmission session consists of two time slots (TS). In the first TS, the EN transmits $x_{2,k}$ to user k while receiving $x_{1,k}$ from the WAP. In the second TS, the EN forwards $x_{1,k}$ to user k. The received signal at the EN, $y_{1,k}$, and at user k in the first

TS, $y_{2,k}^{(1)}$, is given as:

$$y_{1,k} = \sqrt{P_1}\bar{\boldsymbol{h}}\boldsymbol{w}^T x_{1,k} + \sqrt{P_2} h_{SI} x_{2,k} + n_{1,k},$$
$$y_{2,k}^{(1)} = \sqrt{P_2}\bar{g}_k x_{2,k} + n_{2,k}^{(1)},$$

where P_1, P_2 are the transmit power at the WAP and the EN, respectively, $\boldsymbol{w}$ is the precoding vector for the backhaul link, h_{SI} is the self-interference channel at the EN, $n_{1,k}$ and $n_{2,k}^{(1)}$ are Gaussian noises with zero-mean and variance σ^2.

In order to decode for $x_{1,k}$, the EN performs interference cancellation since $x_{2,k}$ is known. After interference cancellation, the residual interference power is ηP_2, where η is a constant[1] depending on the cancellation efficiency [18, 19]. In the second TS, the EN forwards $x_{1,k}$ to the user. The received signal at user k in the second TS is $y_{2,k}^{(2)} = \sqrt{P_2}\bar{g}_k x_{1,k} + n_{2,k}^{(2)}$. The effective achievable rate for user k on the backhaul link and access link are given, respectively, as

$$R_{1,k} = \frac{B}{K}\log_2\left(1 + \frac{P_1 G_1 |\boldsymbol{h}\boldsymbol{w}^T|^2}{\eta P_2 + \sigma^2}\right),$$
$$R_{2,k} = \frac{B}{K}\log_2\left(1 + \frac{P_2 G_{2,k} |g_k|^2}{\sigma^2}\right).$$

3 Delivery time analysis

In this section, we analyse the average delivery time for the requested content over the fading channels. Because the library size is usually large compared to the number of users, it is highly probable that the users request different contents. Since the analysis for the users are similar, we drop the user subscript k for simplicity and use R_1, R_2, G_2 instead of $R_{1,k}, R_{2,k}, G_{2,k}$, respectively. The average delivery time is calculated as

$$T = \sum\nolimits_{n=1}^{N} \bar{T}_n q_n, \tag{2}$$

[1] Considering a random model of η is left for future work.

where q_n is given in (1) and $\bar{T}_n$ is the average delivery time (over the fading channels) when requesting file F_n, which is computed as follows:

$$\bar{T}_n = \mathbb{E}_{\boldsymbol{h},g}[T_n(\boldsymbol{h}, g)], \tag{3}$$

where $T_n(\boldsymbol{h}, g)$ is the (instantaneous) delivery time of file F_n, which depends on the instantaneous channel fading coefficients.

The instantaneous delivery time consists of two parts corresponding to two TSs discussed in Section 2.2. Note that c_nQ bits (corresponding to signal $x_{2,k}$) of the requested file are available at the EN, and the BS needs to send the remaining $(1-c_n)Q$ bits (corresponding to signal $x_{1,k}$). By employing the FD mode, the EN transmits c_nQ bits to the user while receiving $(1-c_n)Q$ bits from the WAP during the first TS. Therefore, the instantaneous delivery time is calculated as

$$\begin{aligned} T_n(\boldsymbol{h}, g) &= \max\left(\frac{(1-c_n)Q}{R_1}, \frac{c_nQ}{R_2}\right) + \frac{(1-c_n)Q}{R_2} \\ &= T_n^{(1)}(\gamma_1, \gamma_2) + T_n^{(2)}(\gamma_2), \end{aligned} \tag{4}$$

where $T_n^{(1)}(\gamma_1, \gamma_2) \triangleq \max\left(\frac{(1-c_n)QK}{B\log_2(1+\gamma_1)}, \frac{c_nQK}{B\log_2(1+\gamma_2)}\right)$, $T_n^{(2)}(\gamma_2) \triangleq \frac{(1-c_n)QK}{B\log_2(1+\gamma_2)}$, $\gamma_1 \triangleq \frac{P_1G_1|\boldsymbol{h}\boldsymbol{w}^T|^2}{\eta P_2+\sigma^2}$ and $\gamma_2 \triangleq \frac{P_2G_2|g|^2}{\sigma^2}$.

Define $\gamma_2^\star \triangleq (1+\gamma_1)^{\frac{c_n}{1-c_n}} - 1$, we can further express the delivery time $T_n^{(1)}(\gamma_1, \gamma_2)$ as follows:

$$T_n^{(1)}(\gamma_1, \gamma_2) = \begin{cases} \frac{(1-c_n)QK}{B\log(1+\gamma_1)}, \text{ if } \gamma_2 \geq \gamma_2^\star \\ \\ \frac{c_nQK}{B\log(1+\gamma_2)}, \text{ if } \gamma_2 < \gamma_2^\star \end{cases}. \tag{5}$$

In order to compute $\bar{T}_n$ in (3), we first need to know the probability density function (pdf) of γ_1 and γ_2. Under the Rayleigh fading, it is straightforward to see that γ_2 follows an exponential distribution with the pdf $f_{\gamma_2}(x) = \exp(-x/\bar{\gamma}_2)/\bar{\gamma}_2$, where $\bar{\gamma}_2 = \frac{P_2G_2}{\sigma^2}$. Under the maximum ratio transmission (MRT) precoding, i.e., $\boldsymbol{w} = \boldsymbol{h}^*$, where $()^*$ denotes the conjugate, we can verify that γ_1 follows a Gamma distribution with the pdf $f_{\gamma_1}(x) = \frac{x^{L-1} e^{\frac{x}{\bar{\gamma}_1}}}{\bar{\gamma}_1^L \mathit{\Gamma}(L)}$, where $\bar{\gamma}_1 = P_1 G_1/\sigma^2$ and $\mathit{\Gamma}(x)$ is the Gamma function.

Given the pdf of γ_1, γ_2, we can calculate the average delivery time as

$$\begin{aligned}
\bar{T}_n =& \mathbb{E}_{\boldsymbol{h},g}[T_n(\boldsymbol{h}, g)] \\
=& \mathbb{E}_{\gamma_1,\gamma_2}[T_n^{(1)}(\gamma_1, \gamma_2) + T_n^{(2)}(\gamma_2)] \\
\overset{(a)}{=}& \underbrace{\mathbb{E}_{\gamma_1,\gamma_2}[T_n^{(1)}(\gamma_1, \gamma_2)]}_{\bar{T}_n^{(1)}} + \underbrace{\mathbb{E}_{\gamma_2}[T_n^{(2)}(\gamma_2)]}_{\bar{T}_n^{(2)}},
\end{aligned} \tag{6}$$

where (a) is because $T_n^{(2)}(\gamma_2)$ depends only on γ_2.

By using (5) while noting that $\gamma_2^\star$ is a function of γ_1, we have:

$$\begin{aligned}
\bar{T}_n^{(1)} =& \int_{\gamma_1=\gamma_{\min}}^{+\infty} \Big(\int_{\gamma_{\min}}^{\gamma_2^\star} T_n^{(1)}(\gamma_1, \gamma_2) + \int_{\gamma_2^\star}^{+\infty} T_n^{(1)}(\gamma_1, \gamma_2) \Big) \\
& \qquad \times f_{\gamma_2}(\gamma_2) d\gamma_2 f_{\gamma_1}(\gamma_1) d\gamma_1 \\
=& \int_{\gamma_1=\gamma_{\min}}^{+\infty} \Big(\int_{\gamma_{\min}}^{\gamma_2^\star} \frac{c_n K Q}{B \log(1+\gamma_2)} \frac{e^{-\frac{\gamma_2}{\bar{\gamma}_2}}}{\bar{\gamma}_2} d\gamma_2 \Big) f_{\gamma_1}(\gamma_1) d\gamma_1 \\
+& \int_{\gamma_1=\gamma_{\min}}^{+\infty} \Big(\int_{\gamma_2^\star}^{+\infty} \frac{(1-c_n) Q K}{B \log(1+\gamma_1)} \frac{e^{-\frac{\gamma_2}{\bar{\gamma}_2}}}{\bar{\gamma}_2} d\gamma_2 \Big) f_{\gamma_1}(\gamma_1) d\gamma_1,
\end{aligned} \tag{7}$$

where $\gamma_{\min}$ is the minimum signal to noise ratio that the system can operate reliably (depending on modulation and coding scheme).

After some algebraic manipulations, we can express

$$\bar{T}_n^{(1)} = \int_{\gamma_1=\gamma_{\min}}^{+\infty} \left(\mathit{\Psi}_1(\gamma_1, \gamma_2) + \mathit{\Psi}_2(\gamma_1) \right) d\gamma_1, \tag{8}$$

where

$$\mathit{\Psi}_2(\gamma_1) = \frac{(1-c_n) Q K {\gamma_1}^{L-1} e^{-\frac{1}{\bar{\gamma}_2}(\gamma_1+1)^{\frac{c_n}{1-c_n}} - \frac{\gamma_1}{\bar{\gamma}_1}}}{B \bar{\gamma}_1^L \mathit{\Gamma}(L) \log(1+\gamma_1)},$$

and

$$\mathit{\Psi}_1(\gamma_1, \gamma_2) = \frac{c_n K Q {\gamma_1}^{L-1} e^{-\frac{\gamma_1}{\bar{\gamma}_1}}}{B \bar{\gamma}_1^L \mathit{\Gamma}(L)} \Big(\int_{\gamma_{\min}}^{(1+\gamma_1)^{\frac{c_n}{1-c_n}} - 1} \frac{e^{-\frac{\gamma_2}{\bar{\gamma}_2}}}{\bar{\gamma}_2 \log(1+\gamma_2)} d\gamma_2 \Big).$$

Similarly, we can compute $\bar{T}_n^{(2)}$ as

$$\bar{T}_n^{(2)} = \int_{\gamma_{\min}}^{+\infty} \frac{(1-c_n)QKe^{-\frac{\gamma_2}{\bar{\gamma}_2}}}{\bar{\gamma}_2 B \log(1+\gamma_2)} d\gamma_2. \tag{9}$$

Substituting (8) and (9) into (6) and then into (2), we can obtain the average delivery time by using numerical integration method. Based on the derived formula, impacts of the key system parameters, e.g., transmit power and cache size, can be inferred directly.

4 Delivery time minimization

In this section, we aim at optimizing the transmit power to minimize the delivery time. Since the optimization for one user is independent from others, we drop the user index for simplicity. Denote p_1 as the transmit power of the WAP, and p_2, p_3 as the transmit power of the EN in the first and second time slots, respectively. Then the delivery time[2] is expressed as follows:

$$\begin{aligned} T_n(\boldsymbol{h}, g) = & \max\Big(\frac{K_1}{\log(1+\frac{p_1 G_1 |\boldsymbol{h}\boldsymbol{w}^T|^2}{\eta p_2+\sigma^2})}, \frac{K_2}{\log(1+\frac{p_2 G_2|g|^2}{\sigma^2})}\Big) \\ & + \frac{K_1}{\log(1+\frac{p_3 G_2 |g|^2}{\sigma^2})}, \end{aligned} \tag{10}$$

where $K_1 \triangleq \frac{(1-c_n)QK}{B}$ and $K_2 \triangleq \frac{c_n QK}{B}$.

The delivery time minimization problem can be formulated as follows:

$$\begin{aligned} & \underset{\{p_i>0\}_{i=1}^3}{\texttt{minimize}} \; T_n(\boldsymbol{h}, g) \\ & \texttt{s.t.} \;\; p_1 \le P_1; \; p_2, p_3 \le P_2, \end{aligned} \tag{11}$$

where P_1 and P_2 are the maximum transmit power at the WAP and the EN, respectively.

[2] The computation time is assumed to be negligible.

In order to deal with the max function in $T_n(\boldsymbol{h}, g)$, we introduce an arbitrary positive variable t and equivalently reformulate problem (11) as follows:

$$\underset{t,\{p_i\}_{i=1}^3}{\texttt{minimize}} \quad t + \frac{K_1}{\log(1+\frac{p_3 G_2|g|^2}{\sigma^2})} \tag{12}$$

$$\texttt{s.t.} \quad K_1 \le t \log\Big(1 + \frac{p_1 G_1 |\boldsymbol{h}\boldsymbol{w}^T|^2}{\eta p_2 + \sigma^2}\Big) \tag{12a}$$

$$K_2 \le t \log\Big(1 + \frac{p_2 G_2 |g|^2}{\sigma^2}\Big) \tag{12b}$$

$$p_1 \le P_1;\ p_i \le P_2, i = 2, 3. \tag{12c}$$

The equivalence between (12) and (11) can be verified since the equalities in (12a) and (12b) hold at the optimum. Because p_3 appears only in constraint (12c) and function $\log(x)$ is a monotonically increasing function in its support, it is straightforward to obtain the optimal value $p_3^\star = P_2$. Therefore, problem (12) is equivalent the following problem:

$$\underset{t,p_1,p_2}{\texttt{minimize}} \quad t + \tau^\star \tag{13}$$

$$\texttt{s.t.} \quad (12a); (12b); p_1 \le P_1; p_2 \le P_2,$$

where $\tau^\star = K_1 / \log(1 + \frac{P_2 G_2 |g|^2}{\sigma^2})$.

Solving problem (13) is challenging since constraints (12a) and (12b) are non-convex. To overcome this difficulty, we reformulate these constraints as $(\sigma^2 + \eta p_2)e^{\frac{K_1}{t}} \le \sigma^2 + \eta p_2 + p_1 G_1 |\boldsymbol{h}\boldsymbol{w}^T|^2$ and $e^{\frac{K_2}{t}} \le 1 + \frac{p_2 G_2 |g|^2}{\sigma^2}$, respectively. Thus, problem (13) can be reformulated as follows:

$$\underset{t,p_1,p_2}{\texttt{minimize}} \quad t \tag{14}$$

$$\begin{aligned} \texttt{s.t.} \quad & e^{\frac{K_2}{t}} \le 1 + p_2 G_2 |g|^2/\sigma^2 \\ & (\sigma^2 + \eta p_2) e^{\frac{K_1}{t}} \le \sigma^2 + \eta p_2 + p_1 G_1 |\boldsymbol{h}\boldsymbol{w}^T|^2 \\ & p_1 \le P_1;\ p_2 \le P_2. \end{aligned}$$

Table 1: Iterative Algorithm to solve (14)

```
1.  Initialize $t_{\min}, t_{\max}$, eps and error $= 1$.
2.  While error $>$ eps do
    2.1. compute $t_H = (t_{\max} + t_{\min})/2$
    2.2. Solve the problem (15)
         If (15) is feasible: $t_{\max} = t_H$;  Else: $t_{\min} = t_H$
    2.3. Compute error $= |t_{\max} - t_{\min}|$
```

It is observed that for a given t, all the constraints of problem (14) are convex. Therefore, we propose to solve (14) iteratively by using the bisection method. The steps of the proposed algorithm are given in Table 1. It is worth noting that problem (15) is always feasible.

$$\begin{aligned} \texttt{find} \quad & p_1, p_2, \\ \texttt{s.t.} \quad & e^{\frac{K_2}{t_H}} \leq 1 + \frac{p_2 G_2 |g|^2}{\sigma^2} \\ & (\sigma^2 + \eta p_2) e^{\frac{K_1}{t_H}} \leq \sigma^2 + \eta p_2 + p_1 G_1 |\boldsymbol{h}\boldsymbol{w}^T|^2 \\ & p_1 \leq P_1; p_2 \leq P_2. \end{aligned} \tag{15}$$

Complexity analysis: The complexity of the proposed algorithm in Table 1 is determined by two factors: i) the computation time of solving (15) and ii) the complexity of the bisection search. Since problem (15) is linear, it has a polynomial complexity. Therefore, the overall complexity of the proposed algorithm depends on the bisection search. In general, it takes $\log_2(\frac{t_{\max} - t_{\min}}{\texttt{eps}})$ steps to converge, where eps is the tolerable error in the bisection search. In other words, the proposed iterative algorithm is linearly convergent.

5 Numerical results

This section presents numerical results to verify the effectiveness of our analysis and optimization. The wireless channels are subject to Rayleigh fading. The EN is 200 meters far from the WAP, which results in a pathloss $G_1 = -90$ dB under the B5a pathloss model [20]. The users are independently and uni-

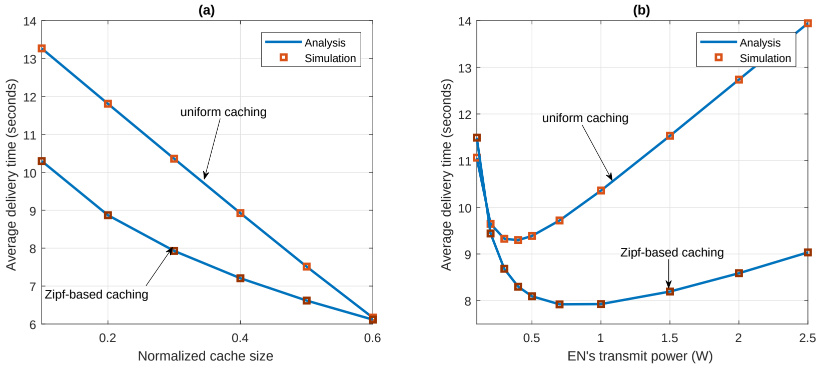

Fig. 2: (a) Delivery time v.s. normalized cache size, $P_2 = 1$ W; (b) Delivery time v.s. P_2, $M = 0.3N$. $\eta = -80$ dB. Solid line shows analytical result, Marker represents simulation result.

formly distributed between 50 to 75 meters from the EN. Therefore, we have -80 dB $\leq G_{2,k} \leq -76$ dB, $\forall k$. In addition, $B = 20$ MHz, $L = 4$, $K = 4$, $P_1 = 20$ W, $\sigma^2 = -110$ dBm, $N = 100$ files, $Q = 100$ Mb, $\gamma_{\min} = 0.07$.

Fig. 2 verifies the accuracy of our analysis for the average delivery time as a function of the normalized cache size $\frac{M}{N}$ (Fig. 2a) and EN's transmit power (Fig. 2b). Two caching strategies are studied: uniform caching, where $\frac{M}{N}$ portions of every file are cached, and Zipf-based caching, where the first M most popular files are cached. The Zipf parameter is 0.8. We observe that the simulation results agree very well with the analysis in all cases. It is also observed that the Zipf-based strategy achieves smaller average delivery times than the uniform caching in both cases. One interesting observation in Fig. 2b is that increasing the EN's transmit power does not always reduce the delivery time. This is because a larger EN's transmit power also creates higher self-interference on the backhaul link. Thus, finding the optimal transmit power is essential in the cache-aided FD system. Because of the superior performance, the Zipf-based caching policy is used in the remaining simulations.

Next, we compare the performance of our proposed optimization scheme with i) reference [13] which applies constant transmit power at the edge node and ii) the half-duplex (HD) which executes backhaul and access transmissions

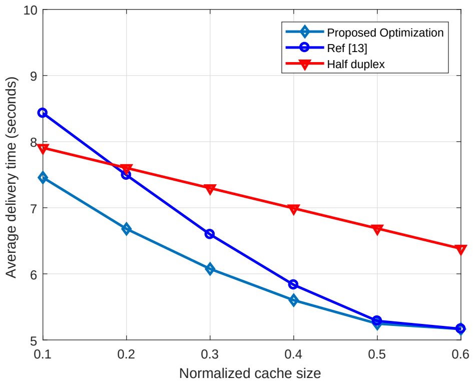

Fig. 3: Performance of the proposed optimization scheme v.s. the normalized cache size, $\eta = -90$ dB and $P_2 = 0.5$ W.

in two orthogonal time slots without interference. The results are obtained over 1000 channel realizations. Fig. 3 presents the delivery time as a function of the normalized cache size, i.e., $\frac{M}{N}$. In general, the proposed optimization significantly outperforms both other schemes. Having a larger cache size results in a smaller delivery time in all schemes since more files can be stored at the edge node. It is also observed that the gain of the proposed optimization over the fixed power transmission is reduced as the cache size increases. In such situation, the required backhaul's rate is less since most of the requested contents are in the EN's cache. Therefore, the EN can operate at a larger transmit power.

Fig. 4 presents the delivery time performance of three schemes as a function of the EN's transmit power P_2. It is observed that the proposed optimization significantly reduces the delivery time compared to [13], specially in the high P_2 regime. This is because the scheme in [13] allows the EN to always transmit at the maximum power, which cause severe self-interference as P_2 increases. Compared to the HD scheme, the proposed optimization algorithm achieves smaller delivery time in the small and medium P_2 regime. When P_2 is large, the HD scheme tends to surpass the FD. This is because at high EN's transmit

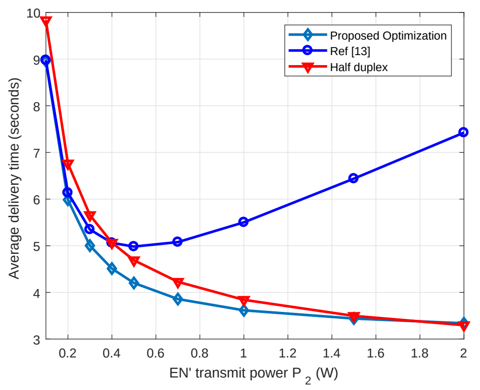

Fig. 4: Performance of the proposed optimization scheme v.s. the EN's transmit power P_2. $\eta = -80$ dB, $M = 0.5N$.

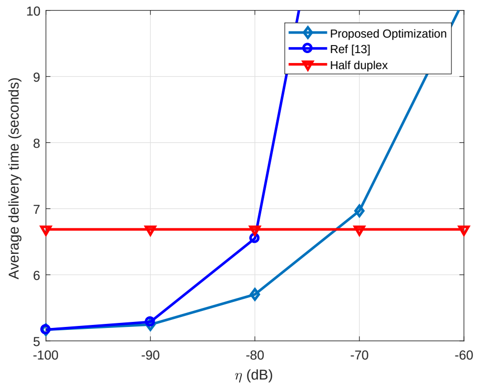

Fig. 5: Performance of the proposed optimization scheme v.s. the interference cancellation efficiency η. $M = 0.5N$ and $P_2 = 0.5$ W.

power, the delivery time on the access link is significantly reduced. In this case, using the HD scheme is preferred.

Fig. 5 compares the delivery time for different interference cancellation efficiency η. Obviously, the HD scheme is independent from η since there is no self-interference. It is observed that the cache-aided FD surpasses the HD scheme when the interference cancellation efficiency is sufficiently small. At the large values regime of η, the FD is limited by the self-interference, hence achieves a higher delivery time than the HD counterpart. As η decreases, the scheme in [13] approaches the optimal performance since the EN can transmit with larger powers without causing severe self-interference on the backhaul.

6 Conclusion

In this paper, we have investigated the performance of cache-aided full-duplex systems via a delivery time metric. In particular, we have derived an analytical expression for the average delivery time over the fading channels, which gives insight to the impact of key system parameters. In addition, we have proposed an power optimization algorithm to minimize the delivery time of the cached-aided full-duplex system.

Based on this result, several extensions are promising for future work. The first topic is to analyze a multiple-antenna setting and imperfect CSI, which will require more advanced precoding and optimization techniques. The second direction is to investigate the on-line caching policy under the FD transmission. In this case, the content popularity will vary over time.

Acknowledgements This work is supported by the Luxembourg National Research Fund under project FNR CORE ProCAST (11691338) and the European Research Council under project AGNOSTIC (742648).

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. 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. T. X. Vu, S. Chatzinotas, B. Ottersten, and A. V. Trinh, "Full-duplex enabled mobile edge caching: From distributed to cooperative caching," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1141–1153, Feb. 2020.
7. L. Yang, Y. Chen, L. Li, and H. Jiang, "Cooperative caching and delivery algorithm based on content access patterns at network edge," *Wireless Networks*, Sep. 2019. [Online]. Available: https://doi.org/10.1007/s11276-019-02148-7
8. J. Ren, T. Hou, H. Wang, H. Ren, and X. Zhang, "Increasing network throughput based on dynamic caching policy at wireless access points," *Wireless Networks*, Aug. 2019. [Online]. Available: https://doi.org/10.1007/s11276-019-02125-0
9. 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.
10. M. E. Ozfatura, S. ElAzzouni, O. Ercetin, and T. ElBatt, "Optimal throughput performance in full-duplex relay assisted cognitive networks," *Wireless Networks*, vol. 25, no. 4, pp. 1931–1947, May 2019. [Online]. Available: https://doi.org/10.1007/s11276-018-1692-5
11. L. Chen, F. R. Yu, H. Ji, B. Rong, and V. C. M. Leung, "Power allocation in small cell networks with full-duplex self-backhauls and massive MIMO," *Wireless Networks*, vol. 24, no. 4, pp. 1083–1098, May 2018. [Online]. Available: https://doi.org/10.1007/s11276-016-1381-1
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.

17. T. X. Vu, L. Lei, S. Chatzinotas, B. Ottersten, and T. A. Vu, “On the successful delivery probability of full-duplex enabled mobile edge caching,” *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 1016–1020, June 2019.

18. 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.

19. 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.

20. M. Hamid and I. Kostanic, “Path loss models for LTE and LTE-A relay stations,” *Universal J. Commun. Netw.*, vol. 1, no. 4, pp. 119–126, 2013.