Energy Minimization for Cache-assisted Content Delivery Networks with Wireless Backhaul

Thang X. Vu, Symeon Chatzinotas, Bjorn Ottersten, and Trung Q. Duong

***Abstract*—Content caching is an efficient technique to reduce delivery latency and system congestion during peak-traffic time by bringing data closer to end users. In this paper, we investigate energy-efficiency performance of cache-assisted content delivery networks with wireless backhaul by taking into account cache capability when designing the signal transmission. We consider multi-layer caching and the performance in cases when both base station (BS) and users are capable of storing content data in their local cache. Specifically, we analyse energy consumption in both backhaul and access links under two uncoded and coded caching strategies. Then two optimization problems are formulated to minimize total energy cost for the two caching strategies while satisfying some given quality of service constraint. We demonstrate via numerical results that the uncoded caching achieves higher energy efficiency than the coded caching in the small user cache size regime.**

***Index terms*—** Edge caching, energy efficiency, wireless backhaul, optimization.

I. Introduction

Edge-caching has recently received much attention as a promising technique to address the stringent requirements of future wireless networks for delivering content at high speed and low latency due to the proliferation of mobile devices and data-hungry applications. The premise of edge-caching is to bring content closer to end users via distributed storage throughout the network [1]. Caching usually consists of two phases: a placement phase which is implemented in off-peak time and a delivery phase which usually occurs during peak-traffic hours when the actual users' requests are revealed. If the requested content is available in the local storage, it can be served locally without being sent via the network. In this manner, caching allows significant throughput reduction during peak-traffic time and thus reducing network congestion [1], [2].

Caching strategies can be classified into two main strategies: *uncoded* [1] and *coded* [2] caching. While uncoded caching strategy prefetches and delivers content to users separately, coded caching requires coordination in both phases to broadcast coded combination to a group of users simultaneously. It is shown in [2] that the coded caching achieves a global caching gain on top of the local caching gain. Note that the global gain brought by coded caching comes at a price of coordination since the data centre needs to know the number of users in order to construct the coded messages. Rate-memory trade-off under coded caching strategy has been intensively investigated in edge-caching networks [2–4] and device-to-device networks [5]. Focusing on the content placement phase in heterogeneous networks, [7] investigates the trade-off between expected backhaul rate and energy consumption. Average energy efficiency (EE) is analysed in [8]. We note that these papers consider only uncoded caching methods with a single cache layer.

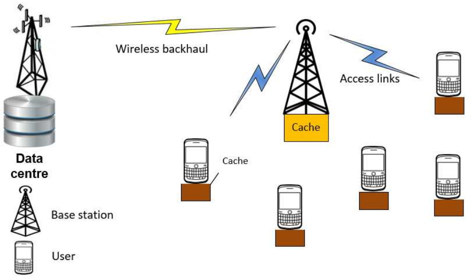


Fig. 1: Cache-assisted content delivery networks.

In this paper, we investigate the energy efficiency performance of cache-assisted content delivery networks (CDN) in which users request content from the data centre via a base station (BS) that is connected to the data centre via wireless backhaul. First, we analyse the energy consumption of the two caching strategies taking into account the wireless channels of the backhaul and access links. We then formulate two optimization problems to minimize the total energy consumption while satisfying a predefined quality of service (QoS). Finally, we demonstrate via numerical results that the uncoded caching achieves higher energy efficiency than the coded caching in the small user cache size regime. Our contributions differ from the existing works in [6–10] since we consider multiple-layer cache, cross-file coded caching and realistic energy cost model for backhaul links.

Notation: $(.)^H, (x)^+$ and $\mathrm{Tr}(.)$ denote the Hermitian transpose, $\max(0,x)$ and the $\mathrm{trace}(.)$ function, respectively. $\lfloor x \rfloor$ denotes the largest integer not exceeding x.

II. Transmission and Caching Model

We consider the downlink cache-assisted CND with wireless backhaul as depicted Fig. 1, in which one BS serving K single-antenna users, denoted by $\mathcal{K} = \{1, \dots, K\}$. The BS connects to the data centre via wireless backhaul which is assumed to operate in a different carrier than the access networks, e.g., mmWave. Therefore, the transmission on the backhaul does not cause interference to the access links. Denote L_1, L_2 with $L_2 \geq K$ as the number of antennas at the data centre and BS, respectively. The wireless transmissions are subjected to block Rayleigh fading channels, in which the channel

Manuscript received Sept. 27, 2017; accepted Nov. 19, 2017. This research is supported, in part, by the ERC AGNOSTIC under code R-AGR-3283 and the FNR ProCAST. The associate editor coordinating the review of this paper and approving it for publication was J. Lee.

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

T. Q. Duong is with the School of Electronics, Electrical Engineering and Computer Science, Queens Belfast University, U.K. (e-mail: trung.q.duong@qub.ac.uk)

fading coefficients are fixed within a block and are mutually independent across links. The block duration is assumed to be sufficiently long to complete a file request section. The data centre contains N files of equal size of Q bits and is denoted by $\mathcal{F} = \{F_1, \dots, F_N\}$.

A. Caching model

We investigate multiple-layer caching networks in which the BS and users are equipped with a storage memory of size M_b and M_u files, with $0 \le M_b, M_u \le N$, respectively. Similar to [2], we consider the worse case in which the caching nodes do not know in advance the content popularity. Two notable caching strategies are considered: uncoded caching and coded caching. In the placement phase under the uncoded caching, content data is preloaded at the caches: BS and each user store $\frac{M_b Q}{N}$ and $\frac{M_u Q}{N}$ bits, which are randomly selected, from every file, respectively. For robustness, we assume the BS is not aware of users' cached content. The placement phase under the coded caching is similar to [2]. In the *delivery phase*, each user requests one file from the data centre. If the requested file parts are in its own cache, they can be served immediately. Otherwise, these parts are sent from the BS's cache or the data centre through the wireless backhaul.

1) Uncoded caching delivery: This strategy sends the parts of the requested file which are not in the local cache to the users separately. The advantage of this method is robustness and it does not require coordination. Denote $Q_{\mathrm{unc,BH}}$ and $Q_{\mathrm{unc,AC}}$ as the aggregated throughput on backhaul and access links in the uncoded caching strategy. Since each user requests a different file, we obtain $Q_{\mathrm{unc,BH}} = KQ\left(1 - \frac{M_u}{N}\right)\left(1 - \frac{M_b}{N}\right)$ and $Q_{\mathrm{unc,AC}} = KQ\left(1 - \frac{M_u}{N}\right)$.

2) Coded caching delivery: In this method, the data centre first intelligently encodes the requested files and then sends them to the users. We note that the data centre needs to know the number of users in order to construct the coded bits.

Proposition 1 ([11]): Let $m = \lfloor \frac{KM_u}{N} \rfloor \in \mathbb{Z}^+$, and $\delta = \frac{KM_u}{N} - m$ with $0 \le \delta < 1$. Under the coded caching strategy, the aggregated throughput on the access links is given as $Q_{\mathrm{cod,AC}} = (1-\delta)\frac{Q(K-m)}{m+1} + \delta\frac{Q(K-m-1)}{m+2}$, and the backhaul thoughtput is calculated as $Q_{\mathrm{cod,BH}} = (1-\delta)\Big(1 - \left(\frac{M_b}{N}\right)^m\Big)\frac{Q(K-m)}{m+1} + \delta\Big(1 - \left(\frac{M_b}{N}\right)^{m+1}\Big)\frac{Q(K-m-1)}{m+2}$.

B. Transmission model

Let $\mathbf{G} \in \mathbb{C}^{L_1 \times L_2}$ and $\mathbf{h}_k \in \mathbb{C}^{L_2 \times 1}, 1 \le k \le K$ denote the channel fading of the wireless backhaul channel and access channels from the BS to the k-th user, respectively, whose elements are mutually independent and follow a complex Gaussian distribution with zero mean and respective variances σ_g^2 and $\sigma_{h_k}^2$. It is assumed that channel state information (CSI) is available at the transmit side, e.g., the data centre knows $\mathbf{G}$ and the BS knows $\mathbf{h}_k, \forall k$. When a user requests a file, it first checks its own cache. Parts of the requested file which are not in neither the user cache nor BS cache will be sent from the data centre to the BS via the wireless backhaul. Then the BS transmits these parts of the file to the user via the access links. The maximum achievable rate that the backhaul supports is $R_{BH} = B_b \log_2(\det(\mathbf{I}_{L_1} + P_{BH}\mathbf{G}\mathbf{G}^H/\sigma^2))$ bps, where B_b is the wireless backhaul bandwidth, $\mathbf{I}_{L_1}$ is the identity matrix of size $L_1 \times L_1$, P_{BH} is the transmit power on the wireless backhaul, and σ^2 is the noise power.

Transmission on the access links under uncoded caching strategy: Denote $\bar{F}_{d_1}, \dots, \bar{F}_{d_K}$ as parts of the requested files which are not available at the user cache. The BS will send these parts to the users. Denote $\mathbf{w}_k \in \mathbb{C}^{L_2 \times 1}$ as the precoding vector for user k. The signal-to-interference-plus-noise ratio at user k is given as $\mathrm{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{l \ne k}|\mathbf{h}_l^H \mathbf{w}_k|^2 + \sigma^2}$, where σ^2 is the noise power. The information achievable rate of user k is $R_{\mathrm{unc},k} = B_a \log_2(1 + \mathrm{SINR}_k), 1 \le k \le K$, where B_a is the access links' bandwidth. In order for user k to successfully decode the requested file, it must hold that $R_{\mathrm{unc},k} \ge \eta, \forall k$, where η is the QoS rate requirement[1]. The transmit power on the access links under the uncoded caching policy is $P_{\mathrm{unc,AC}} = \sum_{k=1}^{K} \parallel \mathbf{w}_k \parallel^2$.

Transmission on the access links under coded caching strategy: In order to optimize the network resources, physical-layer multicasting is employed to precode the data [12] since the coded caching broadcasts the coded-signal to a subset users $\mathcal{K}' \subset \mathcal{K}$ (see [2] for details). In this method, the BS uses only one precoding vector $\mathbf{w} \in \mathbb{C}^{L_2 \times 1}$ for $\mathcal{K}'$. The achievable rate at user $k \in \mathcal{K}'$ is given as $R_{\mathrm{cod}} = \min_k\{R_{\mathrm{cod},k}\}$, where $R_{\mathrm{cod},k} = B_a \log_2\left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}|^2}{\sigma^2}\right)$. To satisfy the QoS rate η, it must hold that $R_{\mathrm{cod}} \ge \eta$. The transmit power on the access links under the coded caching policy is $P_{\mathrm{cod,AC}} = \parallel \mathbf{w} \parallel^2$.

III. Cost minimization for coded caching strategy

We consider off-line caching, in which the *content placement phase* is executed during off-peak time [2]. As such the energy cost in the placement phase is negligible [8] and we focus on the cost in the delivery phase. The total energy cost under the coded caching strategy is given as $E_{\mathrm{cod},\Sigma} = E_{\mathrm{cod,BH}} + E_{\mathrm{cod,AC}}$, where $E_{\mathrm{cod,BH}}$ and $E_{\mathrm{cod,AC}}$ are energy costs in the backhaul and access links, respectively. Since the coded caching strategy sends $Q_{\mathrm{cod,BH}}$ bits via the backhaul link at the rate R_{BH} and $Q_{\mathrm{cod,AC}}$ bits via the access links with the rate R_{cod}, we have $E_{\mathrm{cod,AC}} = \frac{Q_{\mathrm{cod,AC}}}{R_{\mathrm{cod}}} P_{\mathrm{cod}}$ and $E_{\mathrm{cod,BH}} = \frac{Q_{\mathrm{cod,BH}}}{R_{BH}} P_{BH}$. Note that the BS broadcasts a coded message to a subset of users $\mathcal{K}'$ in the delivery phase [2], therefore the required rate on both the backhaul and access links is η.

The energy minimization problem for coded caching strategy is formulated, $\forall k \in \mathcal{K}' \subset \mathcal{K}$, as follows:

$$\mathcal{Q}_0 : \underset{\mathbf{w} \in \mathbb{C}^{L_2 \times 1}, P_{BH}}{\texttt{Minimize}} \; E_{\mathrm{cod},\Sigma} \;\; \texttt{s.t.} \; R_{\mathrm{cod},k} \ge \eta; R_{\mathrm{BH}} \ge \eta.$$

It is observed that problem $\mathcal{Q}_0$ can be decoupled into two sub-problems $\mathcal{Q}_1, \mathcal{Q}_2$ and thus can be effectively solved via optimizing these two problems independently, where

$$\mathcal{Q}_1 : \underset{P_{BH}}{\texttt{Minimize}} \; \frac{Q_{\mathrm{cod,BH}}}{R_{BH}} P_{BH} \;\; \texttt{s.t.} \;\; R_{BH} \ge \eta, \text{ and}$$

$$\mathcal{Q}_2 : \underset{\mathbf{w} \in \mathbb{C}^{L_2 \times 1}}{\texttt{Minimize}} \; \frac{Q_{\mathrm{cod,AC}}}{R_{\mathrm{cod}}} \parallel \mathbf{w} \parallel^2 \;\; \texttt{s.t.} \; R_{\mathrm{cod},k} \ge \eta, \forall k \in \mathcal{K}'.$$

Solving problem $\mathcal{Q}_1$: In order to maximize the achievable rate of the wireless backhaul, water-filling based power allocation is employed. Denote $\lambda_i, 1 \le i \le L_{\min} \triangleq \min(L_1, L_2)$ as

[1] Extension to different η_k for different users is straightforward.

the i-th ordered eigen-values of $\mathbf{GG}^H$, and μ as the water-filling level. Thus, the total transmit power on the backhaul is $P_{BH} = \sum_{i=1}^{L_{\min}}(\mu - 1/\lambda_i)^+$, and the achievable rate of the backhaul is $R_{BH} = B_b \sum_{i=1}^{L_{\min}}(\log_2(\mu\lambda_i))^+$. Then $\mathcal{Q}_1$ is reformulated as $\mathcal{Q}_1^{'}$:

$$\underset{\mu}{\texttt{Min}} \; \frac{\sum_{i=1}^{L_{min}}(\mu - \frac{1}{\lambda_i})^+}{\sum_{i=1}^{L_{min}}(\log_2(\mu\lambda_i))^+}, \; \texttt{s.t.} \; \sum_{i=1}^{L_{\min}}(\log_2(\mu\lambda_i))^+ \geq \frac{\eta}{B_b}.$$

Let $L_{\text{cod}} = \arg\max_l\{l| \sum_{i=1}^{l} \log_2(\lambda_i/\lambda_l) \leq \frac{\eta}{B_b}\}$. The backhaul's transmit power and achievable rate are given in Theorem 1.

Theorem 1: The water-filling level solution μ_{cod} of problem $\mathcal{Q}_1^{'}$ under the uncoded caching strategy is given as $\mu_{\text{cod}} = 2^{\eta/B_b - \sum_{i=1}^{L_{\text{cod}}}\log_2(\lambda_i)}$. Consequently, $P_{\text{cod,BH}} = \sum_{i=1}^{L_{\text{cod}}}(\mu_{\text{cod}} - \frac{1}{\lambda_i})$ and $R_{BH} = B_b \sum_{i=1}^{L_{\text{cod}}} \log_2(\mu_{\text{cod}}\lambda_i)$.

Proof: Denote $f(\mu)$ as the objective function of $\mathcal{Q}_1^{'}$. Consider $f(\mu)$ in an arbitrary subset $[\frac{1}{\lambda_k}, \frac{1}{\lambda_{k+1}}]$ with $1 \leq k \leq L_{min}$, where we denote $\frac{1}{\lambda_{L_{min}+1}} = +\infty$. Since $\{\lambda_i\}$ is a decreasing sequence, we have $f(\mu) = \frac{\sum_{i=1}^{k}(\mu - \frac{1}{\lambda_i})}{\sum_{i=1}^{k}\log_2(\mu\lambda_i)} = \ln(2)\frac{\mu-a}{\ln(\mu)+b}$, where $a = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\lambda_i}$ and $b = \frac{1}{k}\sum_{i=1}^{k}\ln(\lambda_i)$. Taking the derivative of $f(\mu)$ we obtain $f'(\mu) = \frac{\ln(2)}{(\ln(\mu)+b)^2}(\ln(\mu) + \frac{a}{\mu} + b - 1)$. It is straightforward to verify that $\ln(\mu) + \frac{a}{\mu} + b - 1$ is positive in $[\frac{1}{\lambda_k}, \frac{1}{\lambda_{k+1}}]$, which implies that $f(\mu)$ is an increasing function in $[\frac{1}{\lambda_k}, \frac{1}{\lambda_{k+1}}]$. Therefore, we conclude that $f(\mu)$ is an increasing function in $\mathbb{R}^+$. Because the constraint is monotonically increasing function of μ, $f(\mu)$ achieves the minimal value at the smallest μ in the feasible set, i.e., μ_{cod}. ■

In order to solve $\mathcal{Q}_2$, we introduce new variables $x > 0$, $\mathbf{X} = \mathbf{w}^H\mathbf{w} \in \mathbb{C}^{L_2 \times L_2}$ and denote $\mathbf{A}_k = \mathbf{h}_k^H\mathbf{h}_k$. Then the problem $\mathcal{Q}_2$ is equivalent to

$$\underset{\mathbf{X},x}{\texttt{Minimize}} \; \frac{\text{Tr}(\mathbf{X})}{x}, \;\; \texttt{s.t.} \; R_{\text{cod},k} \geq x, \forall k \in \mathcal{K}'; \; x \geq \eta; \\ \mathbf{X} \succeq \mathbf{0}, \; \text{rank}(\mathbf{X}) = 1. \quad (1)$$

We further introduce a new variable $\alpha > 0$ and denote $y = \log_2(x)$, then (1) can be reformulated as follows:

$$\underset{\mathbf{X},y,\alpha}{\texttt{Minimize}} \; \alpha, \;\; \texttt{s.t.} \; \text{Tr}(\mathbf{X}) \leq \alpha\log_2(y); \mathbf{X} \succeq \mathbf{0}; \quad (2) \\ \text{Tr}(\mathbf{A}_k\mathbf{X}) \geq \sigma^2(y-1), \forall k \in \mathcal{K}'; \text{rank}(\mathbf{X}) = 1.$$

For a given α, all constraints of problem (2) are convex except the last one. This suggests to solve problem (2) via bisection and SDR, whose steps are shown in Table I. We note that the solution of SDR does not always satisfy the rank-one condition. Thus, Gaussian randomization procedure is used to obtain the approximated vector from the SDR solution [14]. From the optimal value $\mathbf{X}^\star$ of problem (1), we obtain the optimal precoding vector $\mathbf{w}^\star$ and consequently $P_{\text{cod,AC}}$.

$$\texttt{find} \;\; \mathbf{X} \in \mathbb{C}^{L \times L}, y > 0, \quad \texttt{s.t.} \; \text{Tr}(\mathbf{X}) \leq A_M\log_2(y); \quad (3) \\ \text{Tr}(\mathbf{A}_k\mathbf{X}) \geq \sigma^2(y-1), \forall k \in \mathcal{K}'; \mathbf{X} \succeq \mathbf{0}.$$

TABLE I: ALGORITHM TO SOLVE (1)

```
1. Initialize $A_H$, $A_L = \gamma$, and the accuracy $\epsilon$.
2. $A_M = (A_H + A_L)/2$.
3. Given $A_M$, if (3) is feasible, then $A_H := A_M$.
   Otherwise $A_L := A_M$.
4. Repeat step 2 and 3 until $|A_H - A_L| \leq \epsilon$.
```

IV. COST MINIMIZATION FOR UNCODED CACHING

The total energy cost under the uncoded caching policy is given as $E_{\text{unc},\Sigma} = E_{\text{unc,BH}} + E_{\text{unc,AC}}$, where $E_{\text{unc,BH}}$ and $E_{\text{unc,AC}}$ are the energy cost on the backhaul and access links, respectively. To compute the energy consumption on the access links, we should note that each user requests $\frac{Q_{\text{unc,AC}}}{K}$ bits. The uncoded caching strategy sends these bits to each user independently via multiuser precoding. Since user k is served at the rate $R_{\text{unc},k}$, the energy cost to send the requested content to user k is $\frac{Q_{\text{unc,AC}}}{KR_{\text{unc},k}} \parallel \mathbf{w}_k \parallel^2$. Therefore, the total energy consumed on the access links is $E_{\text{unc,AC}} = \frac{Q_{\text{unc,AC}}}{K}\sum_{k=1}^{K}\frac{\|\mathbf{w}_k\|^2}{R_{\text{unc},k}}$. Since the backhaul needs to support the data rate $K\eta$ for the user request rate η, the energy cost on the backhaul is $\frac{Q_{\text{unc,BH}}}{R_{BH}}P_{BH}$ to transmit the requested bits to the BS. The total energy cost in the uncoded caching strategy is given as

$$E_{\text{unc},\Sigma} = Q(1 - \frac{M_u}{N})\Big(\sum_{k=1}^{K}\frac{\parallel \mathbf{w}_k \parallel^2}{R_{\text{unc},k}} + (1 - \frac{M_b}{N})\frac{P_{BH}}{R_{\text{BH}}}\Big). \quad (4)$$

The energy minimization problem for uncoded caching strategy is formulated as follows:

$$\mathcal{P}_0 : \underset{\mathbf{w}_{k=1:K}, P_{BH}}{\texttt{Minimize}} \, E_{\text{unc},\Sigma}, \; \texttt{s.t.} \; R_{\text{unc},k} \geq \eta, \forall k; R_{\text{BH}} \geq K\eta.$$

where R_{BH} and $R_{\text{unc,k}}$ are given in Section II-B.

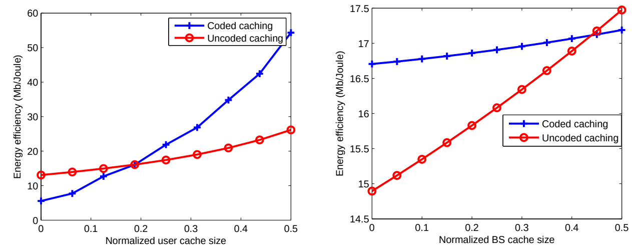
We observe that the problem $\mathcal{P}_0$ can be decoupled because the first constraint depends only on $\mathbf{w}_k$ and the second constraint only affects P_{BH}. Therefore, the solution of $\mathcal{P}_0$ can be obtained by solving two following subproblems: $(\mathcal{P}_1)$ $\underset{P_{BH}}{\texttt{Minimize}} \; \frac{P_{BH}}{R_{\text{BH}}}$ $\texttt{s.t.}$ $R_{\text{BH}} \geq K\eta$ and $(\mathcal{P}_2)$ $\underset{\{\mathbf{w}_k\}_{k=1}^{K}}{\texttt{Minimize}} \; \sum_{k=1}^{K}\frac{\|\mathbf{w}_k\|^2}{R_{\text{unc},k}}$ $\texttt{s.t.}$ $R_{\text{unc},k} \geq \eta, \forall k$.

The solution of problem $\mathcal{P}_1$ can be obtained similarly as problem $\mathcal{Q}_1$ in the previous section. In particular, the backhaul's transmit power under coded caching is given as $P_{\text{unc,BH}} = \sum_{i=1}^{L_{\text{unc}}}(\mu_{\text{unc}} - \frac{1}{\lambda_i})$, where $L_{\text{unc}} = \arg\max_l\{l|\sum_{i=1}^{l}\log_2(\lambda_i/\lambda_l) \leq \frac{K\eta}{B_b}\}$ and $\mu_{\text{unc}} = 2^{\frac{K\eta}{B_b} - \sum_{i=1}^{L_{\text{unc}}}\log_2(\lambda_i)}$.

Solving problem $\mathcal{P}_2$: Due to low computational complexity, Zero-forcing (ZF) is used to design the beamforming vectors. Since the direction of the beamforming vectors are already defined by the ZF, only the transmit power on each beam needs to be optimized. Let $p_k, 1 \leq k \leq K$, denote the transmit power dedicated for user k. The precoding vector for user k is given as $\mathbf{w}_k = \sqrt{p_k}\tilde{\mathbf{h}}_k$, where $\tilde{\mathbf{h}}_k$ is the ZF beamforming vector for user k, which is the k-th column of $\mathbf{H}^H(\mathbf{HH}^H)^{-1}$, with $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]^T$.

Theorem 2: Under the ZF design, the uncoded caching strategy achieves the minimum access energy cost $E_{\text{unc,AC}} = \frac{\zeta\sigma^2\sum_{k=1}^{K}\|\tilde{\mathbf{h}}_k\|^2}{\eta}$, where $\zeta = 2^{\eta/B_a} - 1$.

Proof: By definition, $|\mathbf{h}_l^H\mathbf{w}_k|^2 = p_k\delta_{lk}$, where δ_{ij} is the Dirac delta function. Therefore, the constraint in $\mathcal{P}_2$ becomes $\frac{p_k}{\sigma^2} \geq \zeta, \forall k$. Consequently, the cost minimization problem is



(a) (b)

Fig. 2: (a) - EE v.s. normalized user memory, $M_b = 0.3N$. (b) - EE v.s. normalized BS cache size, $M_u = 0.2N$.

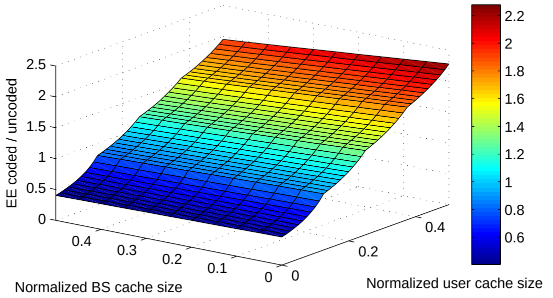


Fig. 3: Relative EE of coded caching v.s. uncoded caching.

formulated as follows:

$$\underset{\{p_k\}_{k=1}^K}{\texttt{Minimize}} \sum_{k=1}^{K} \frac{\|\tilde{\mathbf{h}}_k\|^2 p_k}{\log_2(1+\frac{\|\tilde{\mathbf{h}}_k\|^2}{\sigma^2} p_k)}, \ \texttt{s.t.}\ p_k \geq \zeta\sigma^2, \forall k. \quad (5)$$

Consider a function $f(x) = \frac{ax}{\log_2(1+bx)}$ with $a, b \geq 0$ in $\mathbb{R}^+$. The derivative of $f(x)$ is $f'(x) = \frac{a}{\log_2(1+bx)}(1 - \frac{bx}{\log(1+bx)(1+bx)}) > 0, \forall x > 0$. Therefore, the objective function of (5) is a strictly increasing function in its supports. Therefore, the optimal solution of (5) is achieved at $p_k^\star = \zeta\sigma^2$, and the minimum transmit power is $\zeta\sigma^2 \sum_{k=1}^K \parallel \tilde{\mathbf{h}}_k \parallel^2$. Substituting these into $E_{\text{unc,AC}}$, we obtain Theorem 1. ■

V. Numerical results

This section presents energy efficiency performance of the two caching strategies, which is defined as $EE_{\text{cod/unc}} = \frac{KQ}{E_{\text{cod/unc},\Sigma}}$ measured in bit per Joule. The results are averaged over 1000 channel realizations. $B_b = B_a = 5$ MHz, $K = 8$ users, $L_1 = L_2 = 10$ antennas, $Q = 10$ Mbits, and $\eta = 10$ Mbps. Numerical results are shown for $\frac{M_{u,b}}{N}$ of up to 0.5 since M_u, M_b are usually much smaller than N in practice. Fig. 2a plots the EE v.s. $\frac{M_u}{N}$. Interestingly, there is a crossing point between the two curves, which shows that the uncoded caching strategy outperforms the coded caching for small user cache sizes. This result differs from the common understanding about coded caching that always outperforms the uncoded caching in terms of aggregated backhaul's load [2]. However, [2] ignored the wireless medium between the BS and users. On the other hand, we design the signal transmission carefully for each strategy. In particular, the uncoded caching employs unicast to deliver independent data streams to the users, while the coded caching sends the common coded message to a group of users at a time. When the user cache M_u is greater than $0.2N$, the coded caching achieves higher EE. Fig. 2b shows the EE as a function of BS cache size when $M_u = 0.2N$. In this case, the coded caching achieves higher EE than the uncoded caching at small BS memories. In creasing the BS cache size does not affect much the EE for both strategies.

Fig. 3 presents the ratio of EE of coded caching and EE of uncoded caching as a function of both BS and user cache sizes. A ratio greater than 1 indicates that the coded caching outperforms the uncoded caching method. It is shown that the uncoded caching surpasses the coded caching in the small user cache size regime. When the user cache size is greater than 20% of the library size, the coded caching always outperforms the uncoded caching regardless of BS cache size. This suggests an important guideline for using the uncoded caching method because the user cache size is usually small compared with the library in practice. It is also observed that the user cache size has strong influence on the relative EE gain, whereas the BS cache size has negligible impact.

VI. Conclusions

We have analysed the energy consumption of a cache-assisted CDN with wireless backhaul under the two notable coded and uncoded caching strategies. Two optimization problems have been proposed to minimize the total energy cost for the corresponding two caching methods while satisfying the predefined quality of service requirement. A promising extension based on this work is to consider generic networks in which the data centre is serving multiple base stations and non-identical user cache sizes.

References

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

[2] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Info. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[3] L. Tang and A. Ramamoorthy, "Coded caching for networks with the resolvability property," in *Proc. IEEE ISIT*, Jul. 2016, pp. 420–424.

[4] A. Sengupta, R. Tandon, and T. C. Clancy, "Fundamental limits of caching with secure delivery," *IEEE Trans. Info. Forensics Security*, vol. 10, no. 2, pp. 355–370, Feb. 2015.

[5] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Info. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[6] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud ran," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 6118–6131, Jun. 2016.

[7] F. Gabry, V. Bioglio, and I. Land, "On energy-efficient edge caching in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. PP, no. 99, pp. 3288–3298, Dec. 2016.

[8] 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, 2016.

[9] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Edge-Caching Wireless Networks: Energy-efficient design and optimization," arXiv:1705.05590v2.

[10] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud Radio Access Networks With Coded Caching," in *Proc. Int. ITG WSA*, Munich, Mar. 2016, pp. 1-5.

[11] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Energy-efficient design for edge-caching wireless networks: When is coded-caching beneficial?," in *Proc. IEEE SPAWC*, Sapporo, Jul. 2017.

[12] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[14] 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.