Energy-efficient Design for Edge-Caching Wireless Networks: When is Coded-caching beneficial?

Thang X. Vu, Symeon Chatzinotas, and Bjorn Ottersten
Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg
Email: {thang.vu, symeon.chatzinotas, bjorn.ottersten}@uni.lu

Abstract—**Content caching is an efficient technique to reduce delivery latency and system congestion during peak-traffic times by bringing data closer to end users. Existing works consider caching only at higher layers separated from physical layer. In this paper, we study wireless caching networks by taking into account cache capability when designing the signal transmission. In particular, we investigate multi-layer caching and their performance in edge-caching wireless networks where both base station (BS) and users are capable of storing content data in their local cache. Two notable uncoded and coded caching strategies are studied. Firstly, we propose a coded caching strategy that is applied to arbitrary value of cache size. The required backhaul and access rates are given as a function of the BS and user cache size. Secondly, closed-form expressions for the system energy efficiency (EE) corresponding to the two caching methods are derived. Thirdly, the system EE is maximized via precoding vectors design and optimization while satisfying the user request rate. Finally, numerical results are presented to verify the effectiveness of the two caching methods.**

Index terms— Edge caching, energy efficiency, optimization.

I. Introduction

Future wireless networks will have to address stringent requirements of delivering content at high speed and low latency due to the proliferation of mobile devices and data-hungry applications. It is predicted that by 2020, more than 74% of network traffic will be video [1]. On the other hand, only 5–10% of the files are frequently requested. Various network architectures have been proposed in order to boost the network throughput and reduce transmission latency such as cloud radio access networks and heterogeneous networks. Despite potential high rate in the new architectures, traffic congestion might occur during peak-traffic times. A promising solution to reduce latency and network costs of content delivery is to bring content closer to end users via distributed storage across the network, which is referred to content placement or caching [2] and usually consists of two phases: placement and delivery. In the former, popular content is duplicated and stored in distributed caches in the network. The delivery phase usually occurs during peak-traffic hours when the actual users' requests are revealed. If the requested content is available in the user's local storage, it can be served locally without being sent via the network. In this manner, caching allows significant throughput reduction during peak-traffic times and thus reducing network congestion [2], [3].

Caching strategies can be classified into two main methods: *uncoded* and *coded* caching. The uncoded caching strategy prefetches and delivers content to users independently without coordination [2], [4]. The caching gain can be improved via multicasting a fixed combination of files during the delivery phase in *coded* caching [3]. By carefully prefetching content into the caches and designing coded data, all users can decode their desired content via a multicast stream. It is shown in [3] that the coded caching achieves a global caching gain on top of local caching gain. This gain is inversely proportional to the total cache memory. Rate-memory tradeoff is investigated in device-to-device (D2D) networks [6] and secrecy constraint [5]. In [7], the authors study the tradeoff between the cache memory at edge nodes and the transmission latency. The rate-memory tradeoff of multi-layer coded caching networks is studied in [8]. Note that the global gain brought by the coded caching comes at a price of coordination since the data centre needs to know the number of user in order to construct the coded messages. Recently, energy efficiency (EE) of a cache-assisted networks has been received considerable attentions. Focusing on the content placement phase in heterogeneous networks, the authors in [9] study the trade-off between expected backhaul rate and energy consumption. In [10], the impact of caching is analyzed via close-form expression of the approximated network EE.

In this paper, we investigate the performance of edge-caching wireless networks in which both users and base station (BS) are capable of storing content data [11]. Our contributions are as follows:

- Firstly, we propose a coded-caching strategy that is applied to arbitrary value of cache size. The required rates on backhaul and access links are given.
- Secondly, we analyze the system EE under the two notable caching strategies: uncoded and coded caching. In particular, closed-form expression for the EE is given which reveals insight contributions of cache capability at the BS and users.
- Thirdly, we design and optimize the precoding vectors in order to maximize the system EE while satisfying user request rate. The maximum EE is derived in closed-form for zero-forcing (ZF) precoding and is obtained via semi-definite relaxation (SDR) for general beamforming design. Our paper differs from [9], [10] as following. We focus on the delivery phase, while [10] considers

This work is supported in part by Luxembourg FNR Core program under the project code I2R-SIG-PFN-13SEMI and in part by European Union funding under the project code I2R-SIG-PEU-15SNSA.

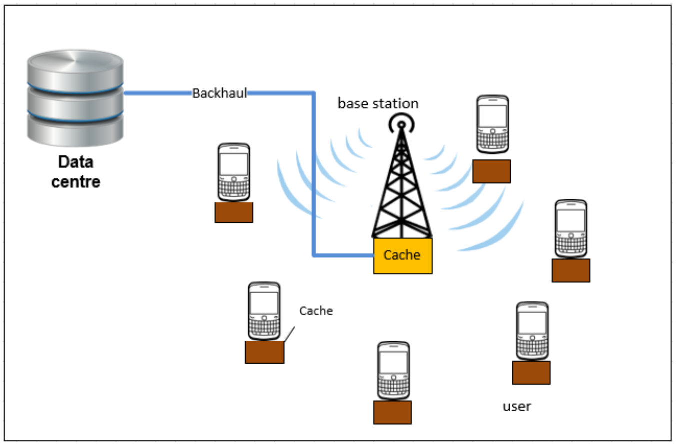


Fig. 1: Cache-assisted wireless networks with heterogeneous user requests and cache capabilities.

the placement phase. We consider multi-layer cache and the two caching strategies, while [9] only consider cache available at BS with uncoded caching algorithm. Finally, the EE is evaluated via numerical results. It is shown that the uncoded-caching is more efficient only for the small user cache sizes.

The rest of this paper is organised as follows. Section II presents the system model and caching strategies. Section III analyses the system EE. Section IV presents the proposed EE maximization algorithms. Section V shows numerical results. Finally, Section VI concludes the paper.

II. System Model

We consider the downlink edge-caching wireless network in which one BS equipped with L antennas serving K single-antenna users with $K \leq L$, as depicted in Figure 1. The BS serves all users via wireless medium and connects to the data centre via a error-free, bandwidth-limited backhaul link. The wireless transmissions are subject to block Rayleigh fading channels, in which the channel fading coefficients are fixed within a block and are mutually independent across the users. The block duration is assumed to be long enough to complete one file request session. The data centre contains N files of equal size of Q bits[1] and is denoted by $\mathcal{F} = \{F_1, \dots, F_N\}$. The library size is NQ bits.

A. Caching model

In order to leverage the traffic in the network, the BS is equipped with a storage memory of size M_b (files) and the users are equipped with a memory of size M_u (files) (equivalent to M_bQ and M_uQ bits), with $0 \leq M_b, M_u \leq N$. We consider off-line caching, in which the *content placement phase* is accomplished during off-peak times [3]. For robustness, we consider the completely distributed placement phase in which the BS is unaware of user cache's content. In particular, the BS randomly stores $\frac{M_bQ}{N}$ bits of every file in its cache, which are randomly chosen. Similarly, each user stores $\frac{M_uQ}{N}$ bits of every file in its cache under the uncoded caching strategy. The placement phase at the user caches under the coded caching is similar to [3].

At the beginning of a file request session, each user requests one file from the data centre. In order to focus on insight interplay between the EE and storage capacity, we assume that the file popularities follow a uniform distribution, i.e., the probability of one file being requested by a user is $\frac{1}{N}$ [3]. The general file popularity distribution is left for future work. Denote $d_1, ..., d_K$ as the file indices requested by user $1, ..., K$, respectively. If the requested file is in the user cache, it can be served immediately. Otherwise, this file is sent from the BS's cache or the data centre through the backhaul link during the *delivery phase*.

We consider two notable caching methods for the delivery phase: uncoded caching and coded caching. We will calculate the required throughput (bits) of the backhaul and access links in order to serve one request session (number of bits sent through backhaul and access links for one file request session).

1) Uncoded caching: The straightforward method is to send the users parts of the file which are not in its local cache [2]. We note that the users do not know the cache content of each other. The advantage of this method is robustness and it does not require coordination. The total number of bits transmitted via the backhaul link, $Q_{\text{unc,BH}}$, and the access links, $Q_{\text{unc,AC}}$, are given in the following proposition.

Proposition 1: Under the uncoded caching strategy, the total number of bits transmitted through the backhaul links is $Q_{\text{unc,BH}} = KQ\left(1 - \frac{M_u}{N}\right)\left(1 - \frac{M_b}{N}\right)$, and the total of bits transmitted through the access links is $Q_{\text{unc,AC}} = KQ\left(1 - \frac{M_u}{N}\right)$.
The proof of Proposition 1 is omitted due to space limitation and can be found by similar techniques in [11, Sec. II].

2) Coded caching: 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 2: Let $m = \lfloor \frac{KM_u}{N} \rfloor \in \mathbb{Z}^+$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x, and $\delta = \frac{KM_u}{N} - m$ with $0 \leq \delta < 1$. Under the coded-caching strategy, the throughput on the access link is $Q_{\text{cod,AC}} = (1-\delta)\frac{Q(K-m)}{m+1} + \delta\frac{Q(K-m-1)}{m+2}$, and the backhaul thoughtput is $Q_{\text{cod,BH}} = (1-\delta)\left(1 - \left(\frac{M_b}{N}\right)^m\right)\frac{Q(K-m)}{m+1} + \delta\left(1 - \left(\frac{M_b}{N}\right)^{m+1}\right)\frac{Q(K-m-1)}{m+2}$.
Proposition 2, where its proof can be found in [14] due to space limitation, derives the access links' throughput under the coded-caching strategy for arbitrary value of M_u. In the special case $\delta = 0$ and $\frac{M_uK}{N} \in \mathbb{Z}^+$, $Q_{cod,AC}$ is shorten as $\frac{KQ(1-M_u/N)}{1+KM_u/N}$, which can also be found in [3]. Note that [3] only derives the access link's rate for limited values of M_u such that $\frac{KM_u}{N}$ is an integer. In other words, Proposition 2 closes the gap in [3] for arbitrary value of the user storage capability M_u.

B. Transmission model

Let $\mathbf{h}_k \in \mathbb{C}^{L\times 1}$ denote the channel vector from the BS antennas to user k, which follows a circular-symmetric complex

[1] In practice, unequal file size can be divided into trunks of subfiles which have the same size.

Gaussian distribution $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \sigma^2_{h_k}\mathbf{I}_K)$, where $\sigma^2_{h_k}$ is the parameter accounting for the path loss from the BS antennas to user k. The BS is assumed to know perfect channel state informations (CSI) from all users. In practice, robust channel estimation can be achieved through the transmission of pilot sequences. When a user requests a file, it first checks its own cache. If the requested file is already in its cache, it can be served immediately. Otherwise, the user sends the requested file's index to the data centre. If the requested file is not at the BS cache, it will be sent from the data centre to send via the backhaul link. Then the BS transmits the requested file to the user via the access links.

1) Signal transmission for uncoded caching strategy: The data stream for each user under the uncoded caching method is transmitted independently. Denote $F_{d_1}, \ldots, F_{d_K}$ as the requested files from user $1, \ldots, K$, respectively, and $\bar{F}_{d_1}, \ldots, \bar{F}_{d_K}$ as parts of the requested files which are not at the user cache. The BS will send these subfiles to the users. First, the BS modulates $\bar{F}_{f_k}$ in to the modulated signal x_k and then sends it through the access channels. The BS first precodes the data before sending to the BS antennas. Denote $\mathbf{w}_k \in \mathbb{C}^{L\times 1}$ as the precoding vector for user k. The received signal at user k is given as $y_k = \mathbf{h}_k^H \mathbf{w}_k x_k + \sum_{l\neq k} \mathbf{h}_k^H \mathbf{w}_l x_l + n_k$, where n_k is Gaussian noise with zero mean and variance σ^2. The first term in y_k is the desired signal, and the second term is the inter-user interference. The signal-to-interference-plus-noise ratio at user k is given as $\mathrm{SINR}_k = \frac{|\mathbf{h}_k^H \mathbf{w}_k|^2}{\sum_{l\neq k} |\mathbf{h}_k^H \mathbf{w}_l|^2 + \sigma^2}$. The information achievable rate of user k is $R_{\mathrm{unc},k} = B\log_2(1 + \mathrm{SINR}_k)$, with $1 \leq k \leq K$ and B is the channel bandwidth. In order for user k successfully decode the requested file, it must hold $R_{\mathrm{unc},k} \geq \gamma, \forall k$, where γ is the minimum request rate.

The transmit power on the access links under the uncoded caching policy is $P_{\mathrm{unc}} = \sum_{k=1}^K \parallel \mathbf{w}_k \parallel^2$.

2) Signal transmission for coded caching strategy: One can use the beamforming design derived in the previous subsection to delivery the requested files in the coded-caching method. However, since the coded caching strategy transmits a common (coded) message $X_{\mathcal{S}}$ (see [3], [14] for details) to all users during the delivery phase, using the orthogonal beams might result in resources redundancy. Instead, the BS applies physical-layer multicasting to precode the coded data [12]. In this method, the BS uses only one precoding vector $\mathbf{w} \in \mathbb{C}^{L\times 1}$. The received signal at user k is given as $y_k = \mathbf{h}_k^H \mathbf{w} x + n_k$, where x is the modulated signal of message $X_{\mathcal{S}}$. The achievable rate at user k under the coded transmission is $R_{\mathrm{cod},k} = B\log_2(1 + \frac{|\mathbf{h}_k^H \mathbf{w}|^2}{\sigma^2})$. In order to successfully deliver the data to all users, it must hold $R_{\mathrm{cod},k} \geq \gamma, \forall k$.

The transmit power on the access links under the coded caching policy is $P_{\mathrm{cod}} = \parallel \mathbf{w} \parallel^2$.

III. Energy-efficiency analysis

This section analyzes the energy-efficiency of the considered networks under two uncoded and coding caching strategies.

Definition 1 (Energy efficiency): The energy efficiency measured in bit/Joule is defined as how many delivered bits per Joule: $\mathrm{EE} = \frac{KQ}{E_\Sigma}$, where KQ is the total requested bits from the K users and E_Σ is the total energy consumption for sending the requested bits in one file request session.

Since the cache placement phase in off-line caching occurs much less frequently (daily or weekly) than the delivery phase, we assume the energy consumed by the placement phase is negligible and thus E_Σ is the energy cost in the delivery phase.

A. EE analysis for uncoded caching strategy

The total energy cost under the uncoded caching policy is given as $E_{\mathrm{unc},\Sigma} = E_{\mathrm{unc,BH}} + E_{\mathrm{unc,AC}}$, where $E_{\mathrm{unc,BH}}$ and $E_{\mathrm{unc,AC}}$ are the energy cost on the backhaul and access links, respectively [13][2]. To compute the energy consumption on the access links, we should note that each user requests $Q_{\mathrm{unc,AC}}/K$ bits. The uncoded caching strategy sends these bits to each user independently via unicasting. Since the user request rate is γ, it takes $\frac{Q_{\mathrm{unc,AC}}}{K\gamma}$ seconds to complete the transmission. Therefore, the total energy consumed on the access links is calculated as $E_{\mathrm{unc,AC}} = \frac{Q_{\mathrm{unc,AC}}}{K\gamma} P_{\mathrm{unc}} = \frac{Q(1-\frac{M_u}{N})}{\gamma} \sum_{k=1}^K \parallel \mathbf{w}_k \parallel^2$.

Sine the backhaul link provides enough capacity to serve the access network, the energy cost on the backhaul is modelled as $E_{\mathrm{unc,BH}} = \eta Q_{\mathrm{unc,BH}} = \eta K Q \left(1 - \frac{M_u}{N}\right)\left(1 - \frac{M_b}{N}\right)$, where η is a constant. In practices, η can be seen as the pricing factor used to trade energy for transferred bits [13]. The actual value of η depends on the backhaul technology.

Therefore, the EE under the uncoded caching strategy is

$$\mathrm{EE}_{\mathrm{unc}} = \frac{K}{\left(1 - \frac{M_u}{N}\right)\left(\eta K\left(1 - \frac{M_b}{N}\right) + \frac{\sum_{k=1}^K \|\mathbf{w}_k\|^2}{\gamma}\right)}. \quad (1)$$

It is observed from (1) that the EE under uncoded caching strategy is jointly determined by the cache capacities M_u and M_b and the transmitted power on the access links.

B. EE analysis for coded caching strategy

The energy cost on the backhaul link under the coded caching strategy is given as $E_{\mathrm{cod,BH}} = \eta Q_{\mathrm{cod,BH}}$, where η is the pricing factor. In order to calculate the energy consumption on the access links, $E_{\mathrm{cod,AC}}$, we note that the BS broadcasts same information $X_{\mathcal{S}}$ to all users in the delivery phase. Therefore, the time duration needed for the BSs to broadcast all $Q_{\mathrm{cod,AC}}$ bits to all the users is $Q_{\mathrm{cod,AC}}/\gamma$. The total energy consumed by the BSs in this case is given as $E_{\mathrm{cod,AC}} = \frac{Q_{\mathrm{cod,AC}}}{\gamma} P_{\mathrm{cod}}$.

The EE in this case is given as $\mathrm{EE}_{\mathrm{cod}} = \frac{KQ}{E_{\mathrm{cod},\Sigma}} = \frac{KQ}{\eta Q_{\mathrm{cod,BH}} + Q_{\mathrm{cod,AC}} \frac{P_{\mathrm{cod}}}{\gamma}}$, where $Q_{\mathrm{cod,AC}}, Q_{\mathrm{cod,BH}}$ are given in Proposition 2. When the user cache size satisfies $\frac{KM_u}{N} \in \mathbb{Z}^+$, i.e., $\delta = 0$, the EE under the coded-caching strategy is shorten as $\mathrm{EE}_{\mathrm{cod}} = \frac{1 + \frac{KM_u}{N}}{\left(1 - \frac{M_u}{N}\right)\left(\eta\left(1 - \left(\frac{M_b}{N}\right)^{\frac{KM_u}{N}+1}\right) + \frac{P_{\mathrm{cod}}}{\gamma}\right)}$.

Similar observation as in the uncoded caching strategy, the EE under the coded-caching is determined by the storage capability at the BS and users and the transmitted power on the access link.

[2] In practice, E_Σ also includes a static energy consumption factor.

IV. Energy-Efficiency Maximization in Cache-assisted wireless networks

We aim at designing the precoding vectors $\{\mathbf{w}_k\}_{k=1}^K$ in order to maximize the EE while satisfying the rate requirement γ. The optimization problem is stated as

$$\underset{\{\mathbf{w}_k\in\mathbb{C}^{L\times 1}\}_{k=1}^K}{\texttt{Maximize}} \quad \text{EE} \qquad \texttt{s.t.}\ R_k \geq \gamma, \forall k, \tag{2}$$

where $\text{EE} \in \{\text{EE}_{\text{cod}}, \text{EE}_{\text{unc}}\}$ stands for the EE under the two caching strategies and $R_k \in \{R_{\text{cod,k}}, R_{\text{unc,k}}\}$, with $R_{\text{unc},k}, R_{\text{cod},k}$ are given in Section II-B. The constraint in (2) is to guarantee reliable transmission on the access links.

A. EE maximization for uncoded caching strategy

It is observed from (1) that for a given network topology, maximizing EE_{unc} is equivalent to minimizing the energy consumption on the access links. Thus, the problem (2) is reformulated as

$$\underset{\{\mathbf{w}_k\}_{k=1}^K}{\texttt{Minimize}} \sum_{k=1}^K \|\,\mathbf{w}_k\,\|^2, \texttt{s.t.} \frac{|\mathbf{h}_k^H\mathbf{w}_k|^2}{\sum_{l\neq k}|\mathbf{h}_k^H\mathbf{w}_l|^2+\sigma^2} \geq \zeta, \forall k. \tag{3}$$

where the rate constraint is replaced by equivalent SINR constraint and $\zeta \triangleq 2^{\frac{\gamma}{B}} - 1$.

1) Cost minimization by Semi-Definite Relaxation: It is observed that problem (3) is a NP-hard problem due to its non-convex constraints. Therefore, we resort (3) into a convex problem for computational efficiency.

We introduce new variables $\mathbf{X}_k = \mathbf{w}_k\mathbf{w}_k^H \in \mathbb{C}^{L\times L}$ and denote $\mathbf{A}_k = \mathbf{h}_k\mathbf{h}_k^H \in \mathbb{C}^{L\times L}$. Taking into consideration the fact that $|\mathbf{h}_l^H\mathbf{w}_k|^2 = \mathbf{h}_l^H\mathbf{w}_k\mathbf{w}_k^H\mathbf{h}_l = \text{Tr}(\mathbf{h}_l\mathbf{h}_l^H\mathbf{w}_k\mathbf{w}_k^H) = \text{Tr}(\mathbf{A}_l\mathbf{X}_k)$, we can reformulate problem (3) as

$$\underset{\{\mathbf{X}_k\in\mathbb{C}^{L\times L}\}_{k=1}^K}{\texttt{Minimize}} \sum_{k=1}^K \text{Tr}(\mathbf{X}_k),\ \texttt{s.t.}\ \mathbf{X}_k \succeq \mathbf{0}; \text{rank}(\mathbf{X}_k) = 1; \tag{4}$$
$$\text{Tr}(\mathbf{A}_k\mathbf{X}_k) \geq \gamma\sum_{l\neq k}\text{Tr}(\mathbf{A}_l\mathbf{X}_k) + \gamma\sigma^2, \forall k.$$

It is observed that problem (4) is still difficult to solve because of the non-convex rank one constraint. Fortunately, the objective function and the two first constraits are convex. Therefore, (4) can be solved by the SDR which is obtained by ignoring the rank one constraint. It can be shown that the SDR of (4) is a convex optimization problem and is solvable by using, *e.g.*, the primal-dual interior point method [15]. From the optimal value $\mathbf{X}_k^\star$ of the SDR of (4), we obtain the optimal precoding vector $\mathbf{w}_k^\star$. Substituting $\mathbf{w}_k^\star$ into (1 we obtain the maximal EE of the uncoded caching strategy under SDR design.

Remark 1: In general, a SDR solution of (4) might violate the rank-one constraint, which is, in fact, a generic problem of SDR. To obtain an approximated (vector) solution $\mathbf{w}_k^*$ for (4) from a SDR counterpart $\mathbf{X}_k^*$, we implement the Gaussian randomization procedure [16].

2) Cost minimization by Zero-Forcing design: Although targeting general beamforming design, the implementation of SDR does not always guarantee the optimum solution, since the rank one constraint might be violated. In this subsection, we maximize the EE based on the ZF design because of its low computational complexity. Since the direction of the beamforming vectors are already defined by the ZF, only transmitting 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 $\mathbf{H}^H(\mathbf{H}\mathbf{H}^H)^{-1}$, with $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]^T$.

Theorem 1: Under the ZF design, the uncoded caching strategy achieves the maximum EE given as

$$\text{EE}_{\text{unc}}^{\text{ZF}} = \frac{K}{\left(1 - \frac{M_u}{N}\right)\left(\eta K\left(1 - \frac{M_b}{N}\right) + \frac{\zeta\sigma^2\sum_{k=1}^K\|\tilde{\mathbf{h}}_k\|^2}{\gamma}\right)}.$$

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 (3) becomes $\frac{p_k}{\sigma^2} \geq \zeta, \forall k$. Consequently, the cost minimization problem is formulated as follows:

$$\underset{\{p_k:p_k\geq 0\}_{k=1}^K}{\texttt{Minimize}} \sum_{k=1}^K p_k \,\|\,\tilde{\mathbf{h}}_k\,\|^2 \quad \texttt{s.t.}\ \ p_k \geq \zeta\sigma^2, \forall k. \tag{5}$$

It is straightforward to verify that the optimal solution of the above problem is $p_k^\star = \zeta\sigma^2$, and the minimum transmit power is $\zeta\sigma^2\sum_{k=1}^K \|\,\tilde{\mathbf{h}}_k\,\|^2$. Substituting this into EE_{unc}, we obtain the proof of Theorem 1. ■

B. EE maximization for coded caching strategy

We observe from (III-B) that maximizing the EE is equivalent to minimizing the transmitted power P_{cod}. The optimization problem in this case is stated as follows:

$$\underset{\{\mathbf{w}\in\mathbb{C}^{L\times 1}\}}{\texttt{Minimize}} \quad \|\,\mathbf{w}\,\|^2 \qquad \texttt{s.t.}\ |\mathbf{h}_k^H\mathbf{w}|^2 \geq \zeta\sigma^2, \forall k. \tag{6}$$

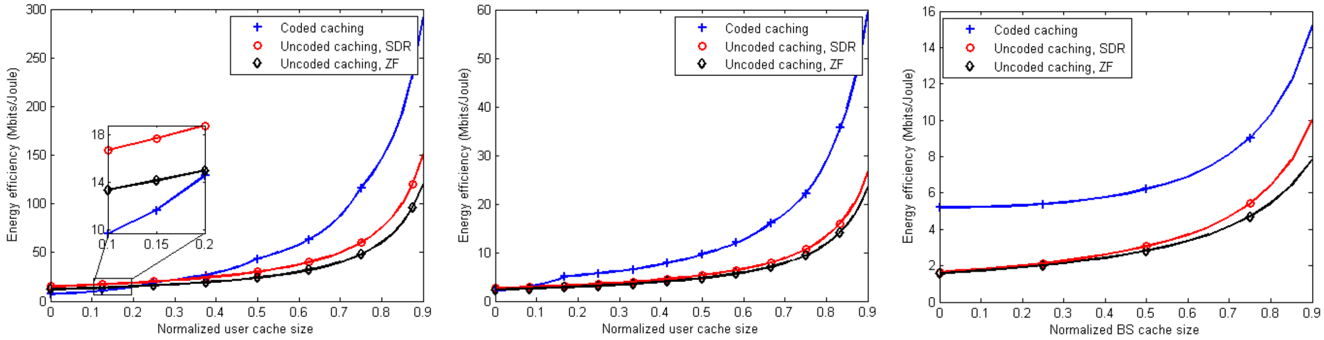
By introducing a new variable $\mathbf{X} = \mathbf{w}^H\mathbf{w} \in \mathbb{C}^{L\times L}$, the problem (6) is equivalent to

$$\underset{\mathbf{X}\in\mathbb{C}^{L\times L}}{\texttt{Minimize}} \ \text{Tr}(\mathbf{X}) \ \ \texttt{s.t.}\ \text{Tr}(\mathbf{A}_k\mathbf{X}) \geq \zeta\sigma^2, \forall k, \tag{7}$$
$$\mathbf{X} \succeq \mathbf{0},\ \text{rank}(\mathbf{X}) = 1.$$

We observe that the objective function and two first constraints of problem (7) are convex. Therefore, by removing the rank-one constraint, the problem (7) can be effectively solved by the SDR method. We note that the solution of SDR does not always satisfy the rank-one condition. Thus, Gaussian randomization procedure might be used to obtain the approximated vector from the SDR solution [16]. From the optimal value $\mathbf{X}^\star$ of problem (7), we obtain the optimal precoding vector $\mathbf{w}^\star$.

V. Numerical results

This section presents numerical results to demonstrate the effectiveness of the studied caching policies. The results are averaged over 300 channel realizations. For ease of presentation, the uncoded caching under the general beamformer design using SDR in Section IV-A1 is named as *SDR* and the ZF design in Section IV-A2 is named as *Zero-Forcing* in the figures.



(a)

(b)

(c)

Fig. 2: a) EE v.s. the normalized user memory, $M_b = N$. b) EE v.s. the normalized user memory, $M_b = 0.5N$. c) EE v.s. the normalized BS memory, $M_u = 0.7N$. In all cases: $L = 10$, $K = 8$, $N = 1000$, $B = 1$ Mhz, $Q = 1$ Mbits, $\gamma = 1$ Mbps, $\sigma^2 = 1$ and $\eta = 10^{-6}$ joule/bit.

We first study the two caching strategies when the energy consumption on the backhaul is negligible. This occurs when the BS cache is large enough to prefetch all the files. In this case, the EE only depends on the user cache size. Figure 2a presents the EE of the two caching strategies as the function of the normalized user cache size (the user cache size M_u divided by the library size N). The EE is plotted based on the optimal precoding vectors obtained from Section IV. It is shown in the figure that the uncoded caching achieves higher EE than the coded caching when the normalized user cache is less than 0.2, which suggests to use uncoded caching when M_u is small. This result is important guideline for practical systems to use the uncoded caching since the user cache is usually capable of storing a small portion of the library in practice. Another observation is that the uncoded caching under SDR design achieves higher EE than the ZF for all user cache size. This is because the SDR design is more efficient than the ZF precoding.

Figure 2b compares the EE for various user cache size when $M_b = 0.7N$. In general, the coded caching method is more efficient than the uncoded caching for all user cache size. Increasing user cache capability results in larger gain of the coded-caching compared with the uncoded method. Figure 2c compares the EE for different BS cache size when $M_u = 0.5N$. Similar conclusion as the previous case that the coded-caching significantly outperforms the uncoded method for all M_b range. It is also shown that the SDR design achieves similar EE as the ZF when M_b is small. From the practical point of view, ZF design is preferred in this case because of its low complexity. When M_b increases, the SDR achieves higher EE than the ZF.

VI. Conclusions

We have analyzed edge-caching wireless networks in which both BS and users are capable of storing content files in their local cache. First, we derived the required rate on backhaul and access links under the coded-caching strategy for arbitrary cache size. Second, we analyzed the system energy efficiency for both coded-caching and uncoded caching strategies. Based on the derived formulas, the EE was maximized via precoding design and optimization. It was shown that the uncoded-caching is more energy efficient than coded caching only when the user cache size is small.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update," 2016-2021 White Paper, 2017.

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

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

[4] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Cellular-broadcast service convergence through caching for COMP cloud RAN," in *Proc. IEEE Symp. Commun. Veh. Tech. in the Benelux*, 2015, pp. 1–6.

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

[6] 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.

[7] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *Proc. Annual Conf. Info. Science and Systems*, Mar. 2016, pp. 320–325.

[8] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical coded caching," *IEEE Trans. Info. Theory*, vol. 62, no. 6, pp. 3212–3229, Jun. 2016.

[9] 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.

[10] 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.

[11] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Coded Caching and Storage Planning in Heterogeneous Networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, San Francisco, CA, Mar. 2017, pp. -

[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, June 2006.

[13] 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.

[14] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Edge-caching wireless networks: Energy-efficient design and optimization," *IEEE Trans. Wireless Commun.*, submitted.

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

[16] 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.