2019 3rd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)

Energy Efficient Design for Coded Caching Delivery Phase

Thang X. Vu*, Lei Lei*, Symeon Chatzinotas*, Björn Ottersten*, and Trinh Anh Vu†

*Interdisciplinary Centre for Security, Reliability and Trust – University of Luxembourg, Luxembourg
†Department of Electronics and Telecommunications, VNU University of Engineering and Technology,
Hanoi, Vietnam. E-Mail: {thang.vu, lei.lei, symeon.chatzinotas, bjorn.ottersten}@uni.lu., vuta@vnu.edu.vn

***Abstract*—Edge-caching is a promising technique to improve the network performance in terms of delivery latency and network congestion during peak-traffic times. Between the two fundamental methods, coded caching has received much attention due to its significant gain over the uncoded counterpart. In this paper, we propose an energy-efficient beamforming design for coded caching delivery phase in wireless networks context. In particular, by exploiting the broadcasting capability of the wireless channels and taking into the cache size, a multi-group multicast based transmission scheme is employed to deliver multiple coded messages to different subgroups of users simultaneously. Numerical results show a significant energy consumption reduction of the proposed design compared to the conventional scheme in the small and medium cache size regime.**

***Index terms*—** coded caching, energy efficiency, multicast, optimization.

I. Introduction

The key challenges of future wireless networks are capable of delivering content at high speed and low latency due to the proliferation of mobile devices and data-hungry applications. Novel network architectures have been proposed in order to boost the network throughput and reduce transmission latency such as cloud radio access networks (C-RANs) [1] and heterogeneous networks (HetNets). Furthermore, it is predicted that by 2020, more than 70% of network traffic will be video [2], and only 5–10% of the files are frequently requested. This unbalanced demands put significant pressure on the backhaul networks, especially during peak hours. Edge caching has received much attention as a promising solution to reduce latency and network costs of content delivery thanks to distributed storages which bring the content closer to end users [3]. In this manner, caching allows significant backhaul's load reduction during peak-traffic times and thus mitigating network congestion [3].

Most research works on caching exploit historic user requested data to optimize either placement or delivery phases [3], [4]. For a fixed content delivery strategy, the placement phase is designed to maximize the local caching gain, which is proportional to the number of file parts available in the local storage. By taking into account the cached content at the edge nodes when designing the signal transmission, caching can bring significant gains in terms of delivery cost and energy efficiency [5]. A joint optimization of caching, routing and channel assignment is proposed in [6] via two restricted master and pricing sub-problems. The stochastic performance of caching wireless networks is analysed in [7] and the impact of node mobility is investigated in [8]. We note that these works consider *uncoded caching* strategy which treats each user request independently.

The caching gain can be further improved via coded caching, which sends a combination of the requested (sub) files to group of users simultaneously during the delivery phase [9], [10]. By carefully placing the files in the caches and designing the coded data, all users can recover their desired content via a multicast stream. It is shown in [9] that the coded caching can achieve a global caching gain additionally to the uncoded caching gain. The rate-memory tradeoff of multi-layer coded caching networks is studied in [11], [12]. The authors in [13], [14] derive an information-theoretic lower bound on the expected transmission rate for arbitrary content popularity. It is worth noting that the benefit of 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. Furthermore, the above mentioned works investigate the coded caching from higher layer aspects separated from the physical layer. In fact, these works focus only on the minimum total transmission rate of the shared backhaul regardless how the requested files are delivered to the users.

Motivated by the above discussion, in this paper, we investigate the coded caching algorithm jointly with the physical layer design and propose an energy-efficient transmission scheme for the coded caching delivery phase. In particular, a multi-group multicast based transmission scheme is employed to send multiple coded messages to different subgroups of users simultaneously thanks to the exploitation of the broadcasting capability of the wireless channels. The idea of using multicast aided coded caching delivery has been used in [15] for computer (wired) systems, and recently applied in wireless networks [16–19]. While the work in [16–19] studied the system from the information-theoretic aspect and assumed perfect superposition decoding with single antenna, we focus on the practical beamforming vectors design and exploit the multiplexing gain of the wireless medium. An optimization problem is formulated to minimize the total energy consumption during the delivery phase while guaranteeing the given quality of service (QoS) constraints. We show via numerical results that the proposed scheme can significantly reduce the total energy consumption.

Notation: $(.)^H$ and $\mathrm{Tr}(.)$ denote the Hermitian transpose the $\mathrm{trace}(.)$ function, respectively. $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$. $\lfloor x \rfloor$ denotes the largest integer not exceeding x. $\binom{n}{k}$ denotes the binomial coefficient.

II. System Model

We consider the cache-assisted wireless network downlink in which a data centre serves K cache-assisted user terminals (UT) [9], [18], denoted by $\mathcal{K} = \{1, \ldots, K\}$, via a base station (BS). The considered system model can also find applications in fog radio access networks or HetNet, where the UTs act the role of small-cell BSs. The BS, equipped with L antennas with $L \geq K$, serves the users' requests via a shared wireless access network. A block Rayleigh fading channel is assumed, in which the channel fading coefficients are fixed within a block and are mutually independent across the links. It is assumed that the block duration is sufficiently long so that the BS can serve the requests within one block [**?**]. The BS is assumed to have full access to the data centre containing a library of N files of equal size of Q bits. The library is denoted as $\mathcal{F} = \{F_1, \ldots, F_N\}$. In practice, unequal-size files can be divided into trunks of subfiles which have same size. Let M denote the cache size (in file) at the ENs. For ease of analysis, we consider $M = m\frac{N}{K}$ for some integers $1 \leq m < K$[1]. We consider off-line caching, in which the *content placement phase* is executed during off-peak times [**?**], [9]. First, each file is divided into $\binom{K}{m}$ subfiles of equal size. Each subfile is of length $Q/\binom{K}{m}$ bits. For convenience, each subfile is associated with a subset of m different UTs in $\mathcal{K}$, i.e., $F_n = \{F_{n,\mathcal{T}} \mid \forall \mathcal{T} \subset \mathcal{K}, |\mathcal{T}| = m\}$. Then in the placement phase, the k-th UT's cache stores $\{F_{n,\mathcal{T}}, \forall n, \mathcal{T} | k \in \mathcal{T}\}$. The details of the placement phase can be found in [9]. The total number of bits stored at the UT caches are MQ bits, which satisf the memory constraint.

In the *delivery phase*, each UT requests one file from the BS. Similarly to [5], [9], we consider the worst case in which the UTs tend to request different files. In coded caching strategy, the data centre first intelligently encodes the requested files and then sends them to the UTs. We note that this strategy requires the number of UTs in order to construct the coded messages for the intended UTs. The total number of bits to be sent through the shared access channel is $\frac{Q(K-m)}{m+1}$ bits.

III. Conventional transmission design

In this section, we describe the conventional transmission design for the delivery phase in coded caching. Let $\mathbf{h}_k \in \mathbb{C}^{L \times 1}$ denote the channel vector from the BS antennas to UT k, which follows a circular-symmetric complex Gaussian distribution $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \sigma^2_{h_k}\mathbf{I}_L)$, where $\sigma^2_{h_k}$ is the parameter accounting for the path loss from the BS antennas to UT k. Perfect channel state information (CSI) is assumed to be available at the BS. In practice, robust channel estimation can be achieved through the transmission of pilot sequences. When a UT requests a file, it first checks its own cache. If (portions of) the requested file is available in its cache, it can be served immediately. Otherwise, the UT sends the requested file's index to the data centre. Then the BS transmits the non-cached parts of the requested file to the user via access links.

In the coded caching strategy, the BS will send $\binom{K}{m+1}$ coded messages (of length $\frac{Q}{\binom{K}{m}}$ bits) in total to the UTs, each of which is received by a subset of $m+1$ UTs [9]. Denote by $\mathcal{S} \subset \mathcal{K}$ an arbitrary subset consisting of $m+1$ UTs, and by $\boldsymbol{\mathcal{S}} = \{\mathcal{S} \mid |\mathcal{S}| = m+1\}$ all subsets of $m+1$ UTs. Obviously, $|\boldsymbol{\mathcal{S}}| = \binom{K}{m+1}$.

Example 1: In the network with $K = 4$ and $M = K/N$, we have $m = 1$. In this case we have 6 subsets of two UTs, i.e., $\boldsymbol{\mathcal{S}} = \{(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)\}$.

Since the coded caching strategy transmits a coded message to a group of UTs during the delivery phase, physical-layer multicasting [20] is used to precode the coded message. For convenience, we denote $X_{\mathcal{S}}$ as the coded message targeted for the UTs in $\mathcal{S}$. The received signal at UT $k \in \mathcal{S}$ is given as $y_k = \mathbf{h}_k^H \mathbf{w}_{\mathcal{S}} x_{\mathcal{S}} + n_k$, where $\mathbf{w}_{\mathcal{S}}$ is the beamforming vector for the UTs in $\mathcal{S}$ and $x_{\mathcal{S}}$ is the unit-power modulated signal of $X_{\mathcal{S}}$, and σ^2 is the noise power. The achievable rate for the UTs in $\mathcal{S}$ under the physical-layer multicasting is [20]

$$R_{\mathrm{con},\mathcal{S}} = \min_{k \in \mathcal{S}} \Big\{ B \log_2 \Big(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_{\mathcal{S}}|^2}{\sigma^2} \Big) \Big\}, \qquad (1)$$

[1] The coded caching scheme for arbitrary cache size, e.g., $M \in (0, N)$, can be obtained in a similar way via the time-split (or memory-sharing) mechanism in which the library is properly divided into two sub-libraries corresponding to cache size mN/K and $(m+1)N/K$, where $m = \lfloor \frac{KM}{N} \rfloor$ [5], [9]

where B is the channel bandwidth. The transmit power on the access links under the coded caching policy is $\| \mathbf{w}_{\mathcal{S}} \|^2$.

Given the QoS constraint, e.g., rate requirement, γ_k, UT k expects to receive the requested file in $t_k = \frac{Q}{\gamma_k}$ seconds. Since each UT receives only $\binom{K-1}{m}$ coded messages out of $\binom{K}{m+1}$, the active time for UT k is $\frac{\binom{K-1}{m}}{\binom{K}{m+1}} t_k = \frac{(m+1)Q}{K\gamma_k}$. Therefore, the required rate for UT k is $\bar{\gamma}_k = (\frac{Q\binom{K-1}{m}}{\binom{K}{m}})/(\frac{(m+1)Q}{K\gamma_k}) = \frac{K-m}{m+1}\gamma_k$, where $\frac{Q\binom{K-1}{m}}{\binom{K}{m}}$ is the total number of coded bits sent to UT k.

With the transmit rate $R_{\text{con},\mathcal{S}}$, the BS is active in $\frac{Q}{R_{\text{con},\mathcal{S}}}$ seconds for sending $x_{\mathcal{S}}$ to all UTs in $\mathcal{S}$. Thus, the energy minimization problem of the conventional design is formulated as [5]:

$$\underset{\mathbf{w}_{\mathcal{S}} \in \mathbb{C}^{L\times 1}}{\mathtt{Minimize}} \; \frac{\| \mathbf{w}_{\mathcal{S}} \|^2}{R_{\text{con},\mathcal{S}}}, \quad \mathtt{s.t.}\ R_{\text{con},\mathcal{S}} \ge \bar{\gamma}_k, \forall k \in \mathcal{S}. \tag{2}$$

where $R_{\text{con},\mathcal{S}}$ is given in (1) and the constraint is to guarantee the rate requirement.

It is worth noting that problem (2) optimizes the beamforming vector for only the UTs in $\mathcal{S}$. Since $\frac{\|\mathbf{w}_{\mathcal{S}}\|^2}{R_{\text{con},\mathcal{S}}}$ is not convex, we resort to finding a suboptimal solution of problem (2) by minimizing the objective's upper bound, i.e., $\frac{\|\mathbf{w}_{\mathcal{S}}\|^2}{R_{\text{con},\mathcal{S}}} \le \frac{\|\mathbf{w}_{\mathcal{S}}\|^2}{\bar{\gamma}_{\min,\mathcal{S}}}$, where $\bar{\gamma}_{\min,\mathcal{S}} = \min_{k\in\mathcal{S}} \bar{\gamma}_k$.

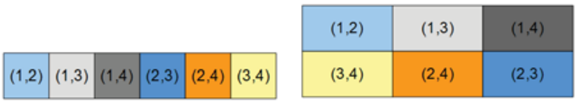
By introducing a new variable $\mathbf{X} = \mathbf{w}_{\mathcal{S}}^H \mathbf{w}_{\mathcal{S}} \in \mathbb{C}^{L\times L}$ and denoting $\mathbf{A}_k = \mathbf{h}_k^H \mathbf{h}_k$, the problem (2) can be reformulated as follows:

$$\begin{aligned} \underset{\mathbf{X} \in \mathbb{C}^{L\times L}}{\mathtt{Minimize}} \; & \frac{\mathrm{Tr}(\mathbf{X})}{\bar{\gamma}_{\min,\mathcal{S}}}, \quad \mathtt{s.t.}\ \mathbf{X} \succeq \mathbf{0};\ \mathrm{rank}(\mathbf{X}) = 1; \\ & \mathrm{Tr}(\mathbf{A}_k \mathbf{X}) \ge \sigma^2 (2^{\bar{\gamma}_{\min,\mathcal{S}}/B} - 1), \forall k \in \mathcal{S}. \end{aligned} \tag{3}$$

By ignoring the rank-one constraint, problem (3) can be solved effectively via semi-definite relaxation (SDR) method [21]. It is noted 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 [21]. From the solution $\mathbf{X}^\star$ of problem (3), we obtain the precoding vector $\mathbf{w}_{\mathcal{S}}^\star$.

IV. Proposed energy efficient design

The conventional transmission design takes advantage from physical-layer multicasting since there is no inter-user interference during the transmission in the delivery phase. In the proposed design, we aim at sending coded messages to multiple subsets of UTs via multi-group multicasting. Although there exists inter-subset interference, the transmit energy is expected to be reduced since the UTs are being served for a larger percentage of time.



(a) Conventional design

(b) Proposed design

Fig. 1: Transmission diagram comparison between the conventional design (a) and the proposed design (b) for a network setup in Example 1. In the conventional design, the BS serves one UT subset, e.g., $(1, 2)$, at a time, whereas the BS in the proposed design serves two subsets simultaneously, e.g., $(1, 2)$ and $(3, 4)$. Each rectangle represents a coded message targets to the UTs within that rectangle.

Denote $v = \lfloor \frac{K}{m+1} \rfloor \in \mathbb{Z}^+$ and let $\boldsymbol{\mathcal{G}}$ denote the collection of v disjoint subsets of $m+1$ UTs, which is defined as

$$\begin{aligned} \boldsymbol{\mathcal{G}} = & \{\mathcal{G}_n \triangleq (\mathcal{S}_{n_1}, \mathcal{S}_{n_2}, \dots, \mathcal{S}_{n_v}) \mid \\ & \mathcal{S}_{n_i} \cap \mathcal{S}_{n_j} = \emptyset, \forall 1 \le i, j \le v\}. \end{aligned}$$

By definition, each $\mathcal{G}_n$ consists of v subsets $\mathcal{S}_{n_i}, 1 \le i \le v$. Consequently, each $\mathcal{G}_n$ contains $(m+1)v$ UTs. For convenience, we name $\mathcal{G}_n$ as the *compound subset*. Since $|\boldsymbol{\mathcal{S}}| = \binom{K}{m+1}$, the cardinality of $\boldsymbol{\mathcal{G}}$ equals to $\lfloor \binom{K}{m+1}/v \rfloor$. The construction of $\boldsymbol{\mathcal{G}}$ can be done via exhausted search. As the result, the original set $\boldsymbol{\mathcal{S}}$ (see details in Section III) is divided into the collection of the compound subsets $\boldsymbol{\mathcal{G}}$ and the remaining subsets $\boldsymbol{\mathcal{S}}_-$ such as $\boldsymbol{\mathcal{S}} = \boldsymbol{\mathcal{G}} \cup \boldsymbol{\mathcal{S}}_-$.

Example 2: In the network with $K = 4$ and $M = K/N$, we have $m = 1$, $v = 2$, and $\boldsymbol{\mathcal{S}} = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$. As the result, we have $\boldsymbol{\mathcal{G}} = \{((1,2),(3,4)),((1,3),(2,4)),((1,4),(2,3))\}$ and $\boldsymbol{\mathcal{S}}_- = \emptyset$.

In the proposed design, the delivery phase is divided into two periods. In the former, the BS multicasts v coded messages to the UTs in one compound subset simultaneously. In the second period, the BS sends one coded message to UTs in a subset $\mathcal{S} \subset \boldsymbol{\mathcal{S}}_-$. The transmission diagram of the proposed design is demonstrated in Fig. 1.

A. Delivery design in the first period

In the first period, the BS serves $(m+1)v$ UTs in the compound subset $\mathcal{G}_n$ (equivalent to v subsets $\mathcal{S}_{n_i}, 1 \le i \le v$) simultaneously. For ease of presentation, we drop the compound subset subscript and use $\mathcal{G}$ as the compound subset of interest. In addition, denote $\mathcal{S}_1, \dots, \mathcal{S}_v$ as the v subsets in the compound subset $\mathcal{G}$. Denote $\mathbf{w}_n, 1 \le n \le v$, as the beamforming vector designed for the UTs in subset $\mathcal{S}_n$. By treating interference as noise, the achievable information rate for the UTs in subset $\mathcal{S}_n, \forall 1 \le n \le v$, is given as

$$R_{\text{prop},\mathcal{S}_n} = \min_{k\in\mathcal{S}_n}\left\{B\log_2\left(1+\frac{|\mathbf{h}_k^H\mathbf{w}_n|^2}{\sum_{n\neq m=1}^{v}|\mathbf{h}_k^H\mathbf{w}_m|^2+\sigma^2}\right)\right\}, \tag{4}$$

where the first term in the denominator represents the inter-subset interference.

Therefore, it takes $Q/R_{\text{prop},\mathcal{S}_n}$ (seconds) for the BS to serve the users in $\mathcal{S}_n$. With the transmit power $\| \mathbf{w}_n \|^2$ for $\mathcal{S}_n$, the total energy consumed to serve all user in the compound subset $\mathcal{G}$ (consists of v subsets $\mathcal{S}_n, n = 1, ..., v$) is $\text{EE} = \sum_{n=1}^{v}\frac{\|\mathbf{w}_n\|^2}{R_{\text{prop},\mathcal{S}_n}}$. Our goal is minimize the energy consumption via proper beamforming vector design of $\mathbf{w}_n, 1 \le n \le v$. The optimization problem is formulated as follows:

$$\begin{aligned} \underset{(\mathbf{w}_n)_{n=1:v}}{\mathtt{Minimize}} \quad & \sum_{n=1}^{v}\frac{\| \mathbf{w}_n \|^2}{R_{\text{prop},\mathcal{S}_n}}, \\ \mathtt{s.t.} \quad & R_{\text{prop},\mathcal{S}_n} \ge \bar{\gamma}_k/v, \forall k \in \mathcal{S}_n, \end{aligned} \tag{5}$$

where $R_{\text{prop},\mathcal{S}_n}$ is given in (4).

It is worth noting that the minimum rate requirement in (5) is v times smaller than the requested rate in problem (2) because the BS is serving v subsets $\mathcal{S}_n$ simultaneously (see Fig. 1 for details).

Finding the exact solution of the above problem is challenging because of the non-convexity of the objective. We instead find a suboptimal solution, by minimizing the upper bound of the objective function. Since $\frac{\|\mathbf{w}_n\|^2}{R_{\text{prop},\mathcal{S}_n}} \le \frac{\|\mathbf{w}_n\|^2}{\bar{\gamma}_{\min,\mathcal{S}_n}}$, where $\bar{\gamma}_{\min,\mathcal{S}_n} = \min_{k\in\mathcal{S}_n}\bar{\gamma}_k$, we have the suboptimal problem written as follows:

$$\begin{aligned} \underset{(\mathbf{w}_n)_{n=1:v}}{\mathtt{Minimize}} \quad & \sum_{n=1}^{v}\frac{\| \mathbf{w}_n \|^2}{\bar{\gamma}_{\min,\mathcal{S}_n}}, \\ \mathtt{s.t.} \quad & R_{\text{prop},\mathcal{S}_n} \ge \frac{\bar{\gamma}_k}{v}, \forall 1 \le n \le v, \forall k \in \mathcal{S}_n. \end{aligned} \tag{6}$$

By introducing a new variable $\mathbf{X}_n = \mathbf{w}_n^H\mathbf{w}_n \in \mathbb{C}^{L\times L}$, the reformulated problem is given as

$$\underset{(\mathbf{X}_n)_{n=1}^{v}}{\mathtt{Minimize}} \quad \sum_{n=1}^{v}\frac{\mathtt{Tr}(\mathbf{X}_n)}{\bar{\gamma}_{\min,\mathcal{S}_n}}, \tag{7}$$

$$\mathtt{s.t.} \quad \mathbf{X}_n \succeq \mathbf{0}, \ \text{rank}(\mathbf{X}_n) = 1, \forall n \tag{7a}$$

$$\frac{1}{2^{\bar{\gamma}_k/v}-1}\mathtt{Tr}(\mathbf{A}_k\mathbf{X}_n) \ge \sum_{m\neq n}\mathtt{Tr}(\mathbf{A}_k\mathbf{X}_m) + \sigma^2,$$

$$\forall m, n \in \{1, \dots, v\}, \forall k \in \mathcal{S}_n. \tag{7b}$$

We note that the constraint above consists of $(m + 1)v$ individual rate constraints for all UTs in $\mathcal{G}$.

It is observed that the objective function and the constraints of problem (3) are convex, except the rank-one constraint. Therefore, SDR method can be employed by ignoring the rank-one constraint. Since the SDR solution does not always satisfy the rank-one condition.

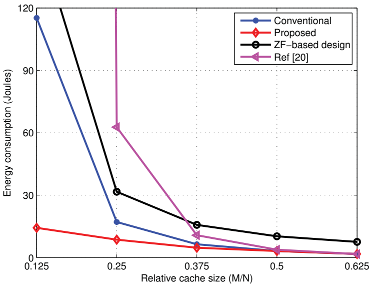


Fig. 2: Energy consumption comparison between the proposed design and the conventional design v.s. the relative cache size $\frac{M}{N}$. The QoS requirement $\gamma_k = 2$ Mbps, $\forall k$.

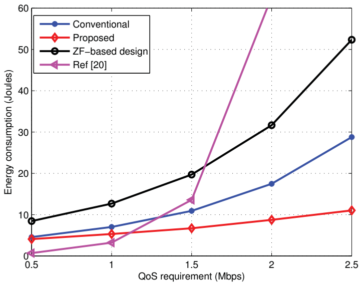


Fig. 3: Energy consumption comparison between the proposed design and the conventional design v.s. the required rate. The cache size $M = 0.25N$.

Thus, Gaussian randomization procedure might be used to obtain the approximated vector from the SDR solution [21]. From the solution $\mathbf{X}^\star$ of problem (3), we obtain the precoding vector $\mathbf{w}_{\mathcal{S}}^\star$.

B. Delivery design in the second period

In the second period, the BS sends a coded message to one subset $\mathcal{S}$ at a time, which is similar to the conventional design in Section III.

Remark 1: When the cache size is large, i.e., $M > \frac{N}{2K}$, then $v = 1$. In this case, it is not possible to do multi-group multicasting. Then the proposed design reduces to the conventional scheme.

V. Numerical results

This section presents numerical results to demonstrate the effectiveness of the proposed transmission design. The results are averaged over 300 channel realizations. Unless otherwise stated, the system parameters are as follows: $L = 10$ antennas, $K = 8$ UTs, $N = 1000$ files, $B = 1$ MHz, $\sigma_{h_k}^2 = 1, \forall k$, $Q = 10$ Mb, and $\sigma^2 = 1$. The proposed design is compared with the conventional scheme [5], [9] described in Section III (named *Conventional* in figure), reference [19], and the zero-forcing based (ZF) design. Since [19] is only applied for single-antenna with superposition coding, the largest antenna coefficient is selected as the channel gain for each user. It is also noted that the ZF design creates orthogonal links among all UTs.

Fig 2 presents the consumed energy of the proposed design and the three references as the function of the relative cache size (the cache size M divided by the library size N). It is observed that the proposed design significantly outperforms the two references when the cache size less than $0.5N$. In particular, at the cache size $M = 0.125N$, the proposed design spends an amount of energy 10 times less than the reference schemes. When the cache size surpasses $0.5N$, the proposed design achieves the same performance as reference [19] and the conventional schemes, as predicted in Section IV. Another observation is that the ZF-based design performs the worst even in the large cache size regime. This is because the ZF design completely mitigates interference among all UTs.

Fig 3 plots the energy consumption for various QoS (required rate) values at the cache size $M = 0.25N$. It is shown that the proposed design always outperforms the ZF and conventional schemes, and the gain increases for a larger required rate. Compared with reference [20], the proposed scheme incurs a higher energy consumption for a small required rate, however, achieves a significant energy reduction as the required rate increases. In this case, the superposition coding scheme is not energy efficient since it spends more energy to suppress the interference.

VI. Conclusions

We have investigated the energy consumption of cache-assisted wireless networks under the coded caching strategy. By exploiting the multicast capability of the wireless channels, we have formulated an optimization problem to minimize the energy consumption during the coded caching delivery phase. It has been shown that the proposed transmission design consumes less energy than the reference schemes in the small and medium cache size regime. The outcome of this work motivates for designing the signal transmission of the coded caching with non-uniform demand in the future.

Acknowledgement

This research is supported, in part, by the Luxembourg National Research Fund under project FNR CORE ProCAST, grant R-AGR-3415-10, and in part by Vietnam National University, Hanoi, under Project No. QG.18.39.

References

[1] T. X. Vu, H. D. Nguyen, T. Q. S. Quek, and S. Sun, “Adaptive cloud radio access networks: compression and optimization,” *IEEE Trans. Signal Process*, vol. 65, no. 1, pp. 228–241, Jan. 2017.

[2] Cisco, “Cisco visual networking index: Global mobile data traffic forecast update 2016-2021,” 2017, white paper.

[3] 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.

[4] K. C. Almeroth and M. H. Ammar, “The use of multicast delivery to provide a scalable and interactive video-on-demand service,” *IEEE J. Sel. Areas Commun.*, vol. 14, no. 6, pp. 1110–1122, IEEE Trans. Inf. Theory. 1996.

[5] T. X. Vu, S. Chatzinotas, and B. Ottersten “Edge-caching Wireless Networks: Performance analysis and optimization,” *IEEE Trans. Wireless Commun.*, vol. pp, no. pp, pp. 1–1, 2018.

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

[7] S. Vuppala, T. X. Vu, S. Gautam, S. Chatzinotas, and B. Ottersten, “Cache-aided millimeter wave Ad-hoc networks,” in *Proc. IEEE Wireless Commun. Netw. Conf.*, Barcelona, Apr. 2018, pp. 1-6.

[8] G. Alfano, M. Garetto, and E. Leonardi, “Content-centric wireless networks with limited buffers: when mobility hurts,” *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 299–311, Jan. 2016.

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

[10] T. X. Vu, S. Chatzinotas, and B. Ottersten, “Coded caching and storage allocation in heterogeneous networks,” in *Proc. IEEE Wireless Commun. Netw. Conf.*, San Francisco, CA, 2017, pp. 1–5.

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

[12] L. Tang and A. Ramamoorthy, “Coded caching for networks with the resolvability property,” in *Proc. IEEE Int. Symp. Inf. Theory*, Barcelona, Jul. 2016, pp. 420–424.

[13] U. Niesen and M. A. Maddah-Ali, “Coded caching with nonuniform demands,” *IEEE Trans. Info. Theory*, vol. 63, no. 2, pp. 1146–1158, Feb. 2017.

[14] J. Zhang, X. Lin, and X. Wang, “Coded caching under arbitrary popularity distributions,” *IEEE Trans. Info. Theory*, vol. 64, no. 1, pp. 349–366, Jan. 2018.

[15] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, “Multi-server coded caching,” *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7253–7271, Dec 2016.

[16] M. A. Maddah-Ali and U. Niesen, “Cache-aided interference channels,” in *Proc. IEEE Int. Symp. Info. Theory*, 2015, pp. 809-813.

[17] J. Zhang and P. Elia, “Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback,” arXiv preprint arXiv: 1511.03961, 2015.

[18] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, “Physical-layer schemes for wireless coded caching,” *CoRR*, vol. abs/1711.05969, 2017. [Online]. Available: http://arxiv.org/abs/1711.05969

[19] M. M. Amiri and D. Gündüz, “Caching and coded delivery over gaussian broadcast channels for energy efficiency,” *CoRR*, vol. abs/1712.03433, 2017. [Online]. Available: http://arxiv.org/abs/1712.03433

[20] 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.

[21] 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.