2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW): Workshop on Intelligent Computing and Caching at the Network Edge

Energy-Efficient Design for Latency-tolerant Content Delivery Networks

Thang X. Vu, Lei Lei, Satyanarayana Vuppala, Symeon Chatzinotas, and Björn Ottersten

The Interdisciplinary Centre for Security, Reliability and Trust (SnT),
University of Luxembourg, 29 Avenue John F. Kennedy, Luxembourg
Email: {thang.vu, lei.lei, satyanarayana.vuppala, symeon.chatzinotas, bjorn.ottersten}@uni.lu

***Abstract*—In this paper, we investigate the energy efficiency performance of content delivery networks in which a data center serves multiple users via a shared wireless medium. Focusing on latency-tolerant applications, we propose energy-efficient precoding design and optimization that minimize the total energy consumption while guaranteeing some given quality of service constraints. In particular, an energy-buffering time trade-off (EBT) is derived in a closed-form expression for single-user scenarios, which reveals the impact of the key system parameters on the total energy consumption. We then formulate an energy minimization problem with a minimum mean square error (MMSE)-based precoding design for multiple-user scenarios. In order to overcome the non-convexity of the formulated problem, we propose an iterative algorithm which solves the problem suboptimally via a linear approximation of the non-convex constraint. Finally, numerical results are presented to demonstrate the effectiveness of the proposed solution.**

***Index terms*—** Content delivery networks, precoding, energy efficiency, latency, optimization.

I. Introduction

Future content delivery networks will have to address stringent requirements of delivering content at high speed and low latency due to the proliferation of mobile handsets and data-hungry applications. It is predicted by Cisco that more than 70% of network traffic will be video in 2018. On the other hand, only 5–10% of the files are frequently requested, which results in an inefficient utilization of network resources of the conventional content delivery. One of the promising solutions to improve the resources utilization is storing the content closer to users in distributed storage, which is referred to content placement or caching [1]. Caching usually consists of two phases: placement and delivery. The placement phase is executed during off-peak time when the network resources are redundant. In this phase, popular content is duplicated and stored in the distributed caches in the network. The later usually occurs during peak-traffic hours when the users' demands are requested. 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 time and thus reduces network congestion [1–5].

The joint design of caching and physical layer design has attracted much attention recently. The basic principle is to take into consideration the caching capacity at the edge nodes when designing the signal transmission to improve the resources [6–9]. The authors in [6] study the trade-off between energy consumption and backhaul load during the placement phase in heterogeneous networks. In [7], a closed-form expression of the energy efficiency is derived showing essential impacts of caching. The authors in [8] show that significant reduction in transmit power and fronthaul bandwidth can be obtained via the careful design of cache-aware multicast beamforming and power allocation. In [10], the authors study D2D networks in which the content can be cached at either small base stations or user nodes. A joint content replacement and delivering scheme is developed to reduce the total energy cost taking into account the fading channels. In [11]], the cache placement design and optimization is investigated for mmWave networks. The authors in [12], [13] study energy consumption based on an over simplified model which assumes caching and transportation costs are linearly dependent on the number of bits. The practical cost model is studied in [14] with wireless backhaul for two caching strategies.

In this paper, we investigate the energy efficiency of content delivery networks in which a base station (BS) is serving multiple users via a shared wireless channel. We focus on latency-tolerant applications where the users can tolerate a reasonable delay before starting the requested service. First, we derive an energy-buffering time trade-off (EBT) in a closed-form expression for single-user scenarios. From the derived closed form, the impact of key system parameters on the total energy consumption is revealed. We then formulate an optimization problem to minimize the total system energy usage for multiple-user scenarios. In order to overcome the non-convexity of the formulated problem, we propose an iterative algorithm which approximates the non-convex constraint by the first order approximation. Finally, the effectiveness of the formulated problem is demonstrated via numerical results.

The rest of this paper is organised as follows. Section II presents the system model. Section III derives the EBT for single-user scenarios. Section IV minimizes the system energy consumption for general multi-user cases. Section V shows numerical results. Finally, Section VI concludes the paper.

II. System Model

We consider a content delivery network consisting of one BS equipped with L antennas serving K single-antenna users via a shared wireless medium, with $K \leq L$, as depicted in

978-1-5386-1154-8/18/$31.00 ©2018 IEEE

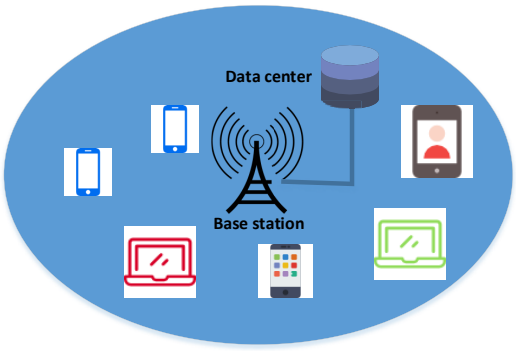

Fig. 1: Content delivery networks via a shared wireless medium.

Figure 1. The BS is connected to a data centre via high speed backhaul links. The BS is assumed to have full access to the content at the data centre, which contains N files of equal size of Q bits (in practice, unequal file size can be divided into trunks of subfiles which have the same size) and is denoted by $\mathcal{F} = \{F_1, \ldots, F_N\}$ the library. The users are equipped with a cache memory of size M (files)[1]. We consider offline caching and focus on the energy consumption of the delivery phase [8].

A. Caching model

In this paper, we assume the content popularity follows a Zipf distribution [15]. The probability of the i-th file being requested from a user is given as

$$f(i) = \frac{i^{-\alpha}}{\sum_{n=1}^{N} n^{-\alpha}}, i = 1, \ldots, N, \quad (1)$$

where α is the skewness factor of the Zipf distribution.

In order to minimize the channel load, the users will cache the most popular files in their cache. In particular, the first M most popular files are prefetched at the user caches during the placement phase, which occurs during off-peak time [1].

B. Signal transmission model

In the delivery phase, each user requests a file from the BS. First the user checks its own cache. If the requested file has been prefetched in its cache, it can be serve immediately. Otherwise, the requested file will be transmitted from the BS. Denote $\mathcal{K}'$ as the subset of users whose requested files are not available in their cache. The BS will only transmit to these users in $|\mathcal{K}'|$. Obviously, $|\mathcal{K}'| \leq K$.

We consider latency-tolerant applications, where the users can allow some buffering time after releasing their requests. Let θ denote a buffering time that the users can tolerate (the gap time between the moment the users send requests and when they can start the requested service, e.g., watching a video). Since the users can tolerate a buffering time θ, they will use this period to preload parts of the requested file to their buffer. Denote $\mathbf{w}_k^b, \mathbf{w}_k^t \in \mathbb{C}^{L\times 1}$ as the precoding vector for user k during the buffering and transmission time, respectively. The received signal at user k is given as

$$y_k^{(b,t)} = \mathbf{h}_k^H \mathbf{w}_k^{(b,t)} x_k + \sum_{k\neq l\in\mathcal{K}'} \mathbf{h}_k^H \mathbf{w}_l^{(b,t)} x_l + z_k, \forall k \in \mathcal{K}' \quad (2)$$

where the superscript (b,t) represents the corresponding buffering time or transmission time, x_k is the modulated signal of the requested file from user k, z_k is Gaussian noise with zero mean and variance σ^2, and $\mathbf{h}_k \in \mathbb{C}^{L\times 1}$ is the channel fading vector from the BS antennas to user k, which follows a circular-symmetric complex Gaussian distribution $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \kappa_k \mathbf{I}_K)$, where κ_k is the parameter accounting for the path loss from the BS antennas to user k. Perfect channel state information (CSI) is assumed to be known at the BS. In practice, robust channel estimation can be achieved through the transmission of pilot sequences. We consider block fading channels and assume the channel coherence time is sufficient long to accommodate one request session [8].

The first term in (2) is the desired signal, and the second term is the inter-user interference. By treating the interference as noise, the respective achievable information rate for user $k \in \mathcal{K}'$ during the buffering time is

$$R_k^b = \mathcal{B}\log_2\left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k^b|^2}{\sum_{l\neq k} |\mathbf{h}_k^H \mathbf{w}_l^b|^2 + \sigma^2}\right),$$

and during the transmission time is

$$R_k^t = \mathcal{B}\log_2\left(1 + \frac{|\mathbf{h}_k^H \mathbf{w}_k^t|^2}{\sum_{l\neq k} |\mathbf{h}_k^H \mathbf{w}_l^t|^2 + \sigma^2}\right),$$

where $\mathcal{B}$ is the channel bandwidth.

Denote r_k as the request rate from user k. With the file length of Q bits, user k expects to receive the requested file in $T_k = \frac{Q}{r_k}$ (seconds). Therefore, to guarantee the smooth experience of the requested sevice, e.g., there is no interruption while watching a movie, the below condition must be hold

$$\theta R_t^b + T_k R_k^t \geq T_k r_k, \forall k \in \mathcal{K}'. \quad (3)$$

III. Delay-tolerant design: single user scenario

In this section, we develop a transmission design for a single-user case. First, we analyze the energy consumption in closed-form and then derive the EBT expression.

A. Minimization of Energy Consumption

For easy of notation, we omit subscript in the rate notation, i.e., r denotes the request rate and omit user index k. In order to maximize the transmission rate, a maximum ratio transmitting precoder is employed. Particularly, the BS applies a precoder $\sqrt{p}\frac{\mathbf{h}^H}{\sqrt{\|\mathbf{h}\|^2}}$ and $\sqrt{q}\frac{\mathbf{h}^H}{\sqrt{\|\mathbf{h}\|^2}}$ during the buffering and transmission time, respectively, where p, q are the corresponding transmit power during the buffering and transmission time. The corresponding achievable rate for the buffering and transmission time is $R_b = \log_2(1+\frac{p\|\mathbf{h}\|^2}{\sigma^2})$ and $R_t = \log_2(1+\frac{q\|\mathbf{h}\|^2}{\sigma^2})$.

[1]Analysis for different cache size, e.g., M_k for user k, is analogous.

We want to minimize the total energy cost $C(\theta) = \theta p + Tq$ to serve the requested file with a tolerable delay θ:

$$\begin{aligned} \underset{\{p,q\}}{\texttt{Minimize}} \quad & \theta p + Tq \qquad (4)\\ \texttt{s.t.} \quad & \theta \log_2(1+\frac{p\|\mathbf{h}\|^2}{\sigma^2}) + T\log_2(1+\frac{q\|\mathbf{h}\|^2}{\sigma^2}) \geq Tr \\ & p \leq P_{tot}; q \leq P_{tot}, \end{aligned}$$

where the first constraint is a reduced form of (3).

Theorem 1: The minimum energy consumption of problem (4) when feasible is

$$C(\mathbf{h}) = \frac{\theta+T}{\|\mathbf{h}\|^2}(2^{\frac{r}{\frac{\theta}{T}+1}}-1)\sigma^2.$$

Proof: Because both the objective function and constraints of problem (4) are convex, duality always holds for the KKT conditions. Consider the Lagrangian function of problem (4) as follows:

$$\begin{aligned} \mathcal{L}(p,q,\lambda_1,\lambda_2,\lambda_3) &= \theta p + Tq \\ &+ \lambda_1(Tr - \theta\log_2(1+\frac{p\,\|\mathbf{h}\|^2}{\sigma^2}) - T\log_2(1+\frac{q\,\|\mathbf{h}\|^2}{\sigma^2})) \\ &+ \lambda_2(p-P_{tot}) + \lambda_3(q-P_{tot}), \end{aligned}$$

where $\lambda_k \geq 0, k = 1,2,3$ is the Lagrangian parameters.

Taking the derivative of $\mathcal{L}(...)$ with respect to its variables, we obtain

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial p} &= \theta - \frac{\theta\lambda_1\,\|\mathbf{h}\|^2}{\ln(2)(\sigma^2+p\,\|\mathbf{h}\|^2)} + \lambda_2 \\ \frac{\partial\mathcal{L}}{\partial q} &= T - \frac{\lambda_1 T\,\|\mathbf{h}\|^2}{\ln(2)(\sigma^2+q\,\|\mathbf{h}\|^2)} + \lambda_3. \end{aligned}$$

By applying the duality and KKT conditions, it yields

$$\theta - \frac{\theta\lambda_1\,\|\mathbf{h}\|^2}{\ln(2)(\sigma^2+p\,\|\mathbf{h}\|^2)} + \lambda_2 = 0 \quad (5)$$

$$T - \frac{\lambda_1 T\,\|\mathbf{h}\|^2}{\ln(2)(\sigma^2+q\,\|\mathbf{h}\|^2)} + \lambda_3 = 0 \quad (6)$$

$$\lambda_1(Tr - \theta\log_2(\sigma^2+p\|\mathbf{h}\|^2) + T\log_2(\sigma^2+q\,\|\mathbf{h}\|^2)) = 0 \quad (7)$$

$$\lambda_2(p-P_{tot}) = 0 \quad (8)$$

$$\lambda_3(q-P_{tot}) = 0. \quad (9)$$

Because the Lagrangian parameters should be non-negative, it must hold $\lambda_1 > 0$. Then there are four cases to consider.

Case 1: $\lambda_2 \neq 0, \lambda_3 \neq 0$. In this case $p = q = P_{tot}$ and the total energy consumption is $E_1 = (\theta+T)P_{tot}$.

Case 2: $\lambda_2 = 0, \lambda_3 = 0$. From (5) and (6) we obtain $p = q$. Then from (7) we have $p = q = \frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{\frac{r}{\theta/T+1}}-1)$. Taking into account the transmit power constraint we finally have $p = q = \min(\frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{\frac{r}{\theta/T+1}}-1), P_{tot})$. The total energy consumption in this case is $E_2 = (\theta+T)\min(\frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{\frac{r}{\theta/T+1}}-1), P_{tot})$.

Case 3: $\lambda_2 \neq 0, \lambda_3 = 0$. From (8) we have $p = P_{tot}$. Substituting $p = P_{tot}$ into (7), it yields $q = \frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{r-\frac{\theta}{T}\log_2(1+P_{tot}\|\mathbf{h}\|^2/\sigma^2)}-1)$. The total energy consumption is $E_3 = \theta P_{tot} + \frac{T\sigma^2}{\|\mathbf{h}\|^2}(2^{r-\frac{\theta}{T}\log_2(1+P_{tot}\|\mathbf{h}\|^2/\sigma^2)}-1)$.

Case 4: $\lambda_2 = 0, \lambda_3 \neq 0$. From (9) we obtain $q = P_{tot}$. From (7), it yields $p = \frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{\frac{r-\log_2(1+P_{tot}\|\mathbf{h}\|^2/\sigma^2)}{(\theta/T)}}-1)$. Therefore, the energy consumption in this case is $E_4 = \frac{\theta}{\|\mathbf{h}\|^2}(2^{\frac{r-\log_2(1+P_{tot}\|\mathbf{h}\|^2/\sigma^2)}{(\theta/T)}}-1) + TP_{tot}$.

Then the optimal solution of Theorem 1 is given as

$$C(\mathbf{h}) = \min\{E_1, E_2, E_3, E_4\}.$$

In order to satisfy the first constraint in (4), the maximum transmit power must satisfy $P_{tot} \geq \frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{\frac{r}{\theta/T+1}}-1)$. Therefore, we have $E_2 = \frac{\theta+T}{\|\mathbf{h}\|^2}(2^{\frac{r}{\theta/T+1}}-1)\sigma^2$, and subsequently $E_1 \geq E_2$. Now consider E_3 as a function of P_{tot}. Its first-order derivative is $E_3'(P_{tot}) = \theta(1-\frac{2^r}{(1+P_{tot}\|\mathbf{h}\|^2/\sigma^2)^{\theta/T+1}}) \geq 0$ as $P_{tot} \geq \frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{\frac{r}{\theta/T+1}}-1)$, which indicates that the function $E_3(x)$ is an increasing function in $[\frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{\frac{r}{\theta/T+1}}-1), +\infty)$. In addition, $E_3(\frac{\sigma^2}{\|\mathbf{h}\|^2}(2^{\frac{r}{\theta/T+1}}-1)) = E_2$. Thus, we have $E_3 \geq E_2$. Similarly, we can verify that $E_4 \geq E_2$. Concluding all cases we obtain $C(\mathbf{h}) = E_2$, which concludes the proof of Theorem 1. ∎

B. Energy-buffering time trade-off

In this section, we analyse the EBT of the single-user scenario, which is defined as the average minimum energy consumption over the fading channels for a given tolerated latency to serve the requesting rate.

Theorem 2 (EBT): With the maximum transmit power P_{tot} and the requested rate r, the EBT is given as

$$\Xi(\theta) = \frac{(\theta+T)\kappa\nu P_{tot}}{(L-1)\Gamma(L)}\left(\Gamma(L;\kappa\nu) - (\kappa\nu)^{L-1}e^{-\kappa\nu}\right),$$

where $\nu = \frac{\sigma^2}{P_{tot}}(2^{\frac{r}{\theta/T+1}}-1)$, $\Gamma(n) = \int_0^{+\infty} x^{n-1}e^{-x}dx$ is the Gamma function, and $\Gamma(n;a) = \int_a^{+\infty} x^{n-1}e^{-x}dx$ is the incomplete Gamma function.

Proof: Denote $\gamma = \|\mathbf{h}\|^2$. Since the elements of $\mathbf{h}$ are i.i.d complex Gaussian random variable with zero mean and variance κ, γ follows the Gamma distribution with the probability density function (pdf) given as

$$f_h(\gamma) = \frac{\kappa^L\gamma^{L-1}e^{-\kappa\gamma}}{\Gamma(L)}.$$

For a given channel realization $\mathbf{h}$, we have the instantaneous energy consumption from Theorem 1 is $E = \frac{(\theta+T)\sigma^2}{\gamma}(2^{\frac{r}{\frac{\theta}{T}+1}}-1)$. We note that in order to guarantee the QoS, it must hold $\gamma \geq \nu \triangleq \frac{\sigma^2}{P_{tot}}(2^{\frac{r}{\theta/T+1}}-1)$. Taking the average over the distribution of γ, we obtain the average energy consumption as follows:

$$\begin{aligned} \mathbb{E}[E] &= \int_\nu^{+\infty} \frac{(\theta+T)\nu P_{tot}}{\gamma} f_h(\gamma)d\gamma \\ &= (\theta+T)\nu P_{tot}\underbrace{\int_\nu^{+\infty}\frac{\kappa^L\gamma^{L-2}e^{-\kappa\gamma}}{\Gamma(L)}d\gamma}_{\mathcal{I}}. \end{aligned} \quad (10)$$

By using partial integration, we obtain:

$$\mathcal{I} = \frac{\kappa^L \gamma^{L-1} e^{-\kappa\gamma}}{(L-1)\Gamma(L)}\Big|_{\nu}^{+\infty} + \int_{\nu}^{+\infty} \frac{\kappa^{L+1}\gamma^{L-1}e^{-\kappa\gamma}}{(L-1)\Gamma(L)} d\gamma$$
$$= \frac{\kappa}{(L-1)\Gamma(L)} \left(\Gamma(L;\kappa\nu) - (\kappa\nu)^{L-1}e^{-\kappa\nu}\right). \quad (11)$$

Substituting (11) into (10) we obtain Theorem 2 proved. ■

From Theorem 2, we can calculate the average EBT over the content popularity. We would note that the BS will transmit to the user only if the requested file is not in the user cache. Given the cache size M, the average energy-buffering time trade-off is computed as:

$$\mathbf{\Xi}(\theta) = \Xi(\theta) \sum_{i=M+1}^{N} f(i), \quad (12)$$

where $f(i)$ is given in (1) and $\Xi(\theta)$ is provided in Theorem 2.

IV. Delay-tolerant design: multiple users scenario

In this section, we analyze the energy consumption in multi-user scenarios. In particular, we want to minimize the total transmit energy for serving all users in $\mathcal{K}'$ (whose requested files are not in their cache), as follows:

$$\underset{\mathbf{w}_k^b, \mathbf{w}_k^t}{\texttt{Minimize}} \quad \sum_{k\in\mathcal{K}'} (\theta \parallel \mathbf{w}_k^b \parallel^2 + T_k \parallel \mathbf{w}_k^t \parallel^2) \quad (13)$$
$$\texttt{s.t.} \quad \theta R_k^b + T_k R_k^t \geq T_k r_k, \forall k \in \mathcal{K}',$$
$$\sum_{k\in\mathcal{K}'} \parallel \mathbf{w}_k^b \parallel^2 \leq P_{tot}; \sum_{k\in\mathcal{K}'} \parallel \mathbf{w}_k^t \parallel^2 \leq P_{tot},$$

where the first constraint in (13) is to guarantee smooth quality of experience when the user start watching the requested file.

We consider two most popular precoding vectors: ZF and MMSE.

A. Zero-Forcing based design

In this subsection, we minimize the energy consumption based on the ZF design because of its low computational complexity. Since the direction of the beamforming vectors are already defined by ZF, only transmitting power on each beam needs to be optimized. The precoding vector for user k is given as $\mathbf{w}_k^b = \sqrt{p_k}\tilde{\mathbf{h}}_k, \mathbf{w}_k^t = \sqrt{q_k}\tilde{\mathbf{h}}_k$, where p_k, q_k is the power factor in the buffering time and transmission time, respectively, $\tilde{\mathbf{h}}_k$ 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 $\mathbf{H} = [\mathbf{h}_{k_1}, \ldots, \mathbf{h}_{k_{|\mathcal{K}'|}}]^T$ denoting the channel matrix from the BS antennas to users in $\mathcal{K}'$. Due to the ZF design, we have $\mathbf{h}_k^H\tilde{\mathbf{h}}_k = 1$ and $\mathbf{h}_k^H\tilde{\mathbf{h}}_l = 0, \forall l \neq k$. Denoting $\alpha_k = \parallel \tilde{\mathbf{h}}_k \parallel^2$, the energy minimization problem is formulated as follows:

$$\underset{\{p_k, q_k\}_{k\in\mathcal{K}'}}{\texttt{Minimize}} \quad \sum_{k\in\mathcal{K}'} \alpha_k(\theta p_k + T_k q_k) \quad (14)$$
$$\texttt{s.t.} \quad \frac{\theta}{T_k}\log_2\left(1+\frac{p_k}{\sigma^2}\right) + \log_2\left(1+\frac{q_k}{\sigma^2}\right) \geq r_k, \forall k \in \mathcal{K}'$$
$$\sum_{k\in\mathcal{K}'} \alpha_k p_k \leq P_{tot}; \sum_{k\in\mathcal{K}'} \alpha_k q_k \leq P_{tot}.$$

We observe that the objective function and the constraints of problem (14) are convex. Thus, it can be solved effectively by, e.g., CVX.

B. MMSE based design

Under MMSE precoding, the beamformer vector is of the form $\mathbf{w}_k^b = \sqrt{p_k}\breve{\mathbf{h}}_k$ during the buffering time and $\mathbf{w}_k^t = \sqrt{q_k}\breve{\mathbf{h}}_k$ during the transmitting time, where $\breve{\mathbf{h}}_k$ is the k-th column of $\mathbf{H}^H(\sigma^2\mathbf{I} + \mathbf{H}\mathbf{H}^H)^{-1}$. Denote $\beta_{k,l} = |\mathbf{h}_k^H\breve{\mathbf{h}}_l|^2, \forall k, l \in \mathcal{K}'$. Then the energy minimization under MMSE design is stated as follows:

$$\underset{\{p_k, q_k\}_{k\in\mathcal{K}'}}{\texttt{Minimize}} \quad \sum_{k\in\mathcal{K}'} \beta_{k,k}(\theta p_k + T_k q_k) \quad (15)$$
$$\texttt{s.t.} \quad \theta \log_2\left(1 + \frac{\beta_{k,k}p_k}{\sum_{k\neq l\in\mathcal{K}'}\beta_{k,l}p_l + \sigma^2}\right) + \quad (15a)$$
$$T_k \log_2\left(1 + \frac{\beta_{k,k}q_k}{\sum_{l\neq k}\beta_{k,l}q_l + \sigma^2}\right) \geq T_k r_k, \forall k \in \mathcal{K}',$$
$$\sum_{k\in\mathcal{K}'} \beta_{k,k}p_k \leq P_{tot}; \sum_{k\in\mathcal{K}'} \beta_{k,k}q_k \leq P_{tot}. \quad (15b)$$

Solving problem (15) is challenging because of the non-convexity of constraint (15a).

First, we denote parameters $A_k = [\sigma^2, \beta_{k,1}, \ldots, \beta_{k,|\mathcal{K}'|}]$, $B_k = [\sigma^2, \beta_{k,1}, \ldots, \beta_{k,k-1}, 0, \beta_{k,k+1}, \ldots, \beta_{k,|\mathcal{K}'|}]$ and introduce new variables $\mathbf{p} = [1, p_{k_1}, \ldots, p_{k_{|\mathcal{K}'|}}]^T$ and $\mathbf{q} = [1, q_{k_1}, \ldots, q_{k_{|\mathcal{K}'|}}]^T$. Furthermore, we introduce new positive variables $\{u_k, v_k\}_{k\in\mathcal{K}'}$. Then the problem (15) is equivalent to

$$\underset{\mathbf{p},\mathbf{q},u_k,v_k}{\texttt{Minimize}} \quad \boldsymbol{\beta}^T\mathbf{p} + \boldsymbol{\beta}^T\mathbf{q} \quad (16)$$
$$\texttt{s.t.} \quad \frac{\theta}{T_k}\log_2(A_k\mathbf{p}) + \log_2(A_k\mathbf{q}) \geq r_k + \frac{\theta}{T_k}u_k + v_k, \forall k \in \mathcal{K}', \quad (16a)$$
$$B_k\mathbf{p} \leq e^{u_k}, \forall k \in \mathcal{K}', \quad (16b)$$
$$B_k\mathbf{q} \leq e^{v_k}, \forall k \in \mathcal{K}', \quad (16c)$$
$$\boldsymbol{\beta}^T\mathbf{p} \leq P_{tot} + \sigma^2; \boldsymbol{\beta}^T\mathbf{q} \leq P_{tot} + \sigma^2, \quad (16d)$$

where $\boldsymbol{\beta} \triangleq [1, \beta_{1,1}, \ldots, \beta_{|\mathcal{K}'|,|\mathcal{K}'|}]^T$.

It is observed that problem (16) is still challenging because the constraints (16b) and (16c) are non-affine. To deal with this, we resort these constraints into linearity by using the first-order Taylor approximation of exponential functions as

$$B_k\mathbf{p} \leq e^{\bar{u}_k}(u_k + 1 - \bar{u}_k)$$
$$B_k\mathbf{q} \leq e^{\bar{v}_k}(v_k + 1 - \bar{v}_k)$$

where $\bar{u}_k, \bar{v}_k$ are arbitrary accessible values.

Because $e^{x_0}(x + 1 - x_0) \leq e^x, \forall x_0$, the approximated constraints give a sub-optimal solutions of the original problem. The resulting problem is as

$$\underset{\mathbf{p},\mathbf{q},u_k,v_k}{\texttt{Minimize}} \quad \boldsymbol{\beta}^T\mathbf{p} + \boldsymbol{\beta}^T\mathbf{q} \quad (17)$$
$$\texttt{s.t.} \quad B_k\mathbf{p} \leq e^{\bar{u}_k}(u_k + 1 - \bar{u}_k), \forall k \in \mathcal{K}', \quad (17a)$$
$$B_k\mathbf{q} \leq e^{\bar{v}_k}(v_k + 1 - \bar{v}_k), \forall k \in \mathcal{K}', \quad (17b)$$

TABLE I: ITERATIVE ALGORITHM TO SOLVE (17)

1. Initialize a_k, b_k, ϵ, $\tau = 1$, P_{old} and `error`.
2. While `error` $> \epsilon$ do
   2.1. Solve $\boldsymbol{\mathcal{P}}_2(\{a_k, b_k\}_{k\in\mathcal{K}'})$ in (18) to obtain optimal values $u_k^\star, v_k^\star, \mathbf{p}^\star, \mathbf{q}^\star$, and $P^{(\tau)} = \text{sum}(\mathbf{p}^\star + \mathbf{q}^\star)$
   2.3. Compute `error` $= |P^{(\tau)} - P_{old}|$
   2.4. Update $P_{old} = P^{(\tau)}, a_k = u_k^\star, b_k = v_k^\star, \tau := \tau + 1$

$(16a)$ and $(16d)$.

We observe that the resorted problem (17) is convex since the objective function and all constraints are convex. Therefore, problem (17) can be effectively solved by, e.g., CVX.

However, the optimal solution of problem (17) heavily relies on parameters $\{\bar{u}_k, \bar{v}_k\}, \forall k$. This raises a question how to choose the values $\{\bar{u}_k, \bar{v}_k\}, \forall k$ such that the suboptimal solution of (17) is as close as possible to the optimal solution of (16). To overcome this problem, we propose an iterative algorithm improve the performance of problem (17), whose steps are listed in Table I.

Proposition 1: The objective function of problem $\boldsymbol{\mathcal{P}}_2$ in (18) solved by the iterative algorithm in Table I decreases by interations.

Proof: Let $\left(\mathbf{p}_\star^{(\tau)}, \mathbf{q}_\star^{(\tau)}, \mathbf{u}_\star^{(\tau)}, \mathbf{v}_\star^{(\tau)}\right)$ be the optimal solution of $\boldsymbol{\mathcal{P}}_\mathbf{2}(\mathbf{a}^{(\tau)}, \mathbf{b}^{(\tau)})$ at the τ-th iteration. The optimal objective function after iteration τ is $P^{(\tau)} = \text{sum}(\mathbf{p}_\star^{(\tau)} + \mathbf{q}_\star^{(\tau)})$. We will show that if either $u_{\star k}^{(\tau)} < a_k^{(\tau)}, \forall k$ or $v_{\star k}^{(\tau)} < b_k^{(\tau)}, \forall k$, then by using $a_k^{(\tau+1)} = u_{\star k}^{(\tau)}$ or $b_k^{(\tau+1)} = v_{\star k}^{(\tau)}$ for the $(\tau+1)$-th iteration, we will have $P^{(\tau+1)} < P^{(\tau)}$. Lets first consider the case $u_{\star k}^{(\tau)} \neq a_k^{(\tau)}, \forall k$. By choosing a relatively large initial value $a_k^{(1)}$, we always have $u_{\star k}^{(1)} < a_k^{(1)}, \forall k$.

At the $(\tau+1)$-th iteration, $f(x; u_{\star k}^{(\tau)})$ is used in the right-hand side of constraint (18b) instead of $f(x; a_k^{(\tau)})$, where $f(x; a) = e^a(x + 1 - a)$ is the first-order approximation of function e^x at a. Consider a candidate $\bar{\mathbf{u}} = \{\bar{u}_1, \ldots, \bar{u}_K\}$, with $\bar{u}_k = u_{\star k}^{(\tau)} - 1 + e^{a_k^{(\tau)} - u_{\star k}^{(\tau)}}(u_{\star k}^{(\tau)} - a_k^{(\tau)} + 1)$. It is straightforward to verify that $\bar{u}_k < u_{\star k}^{(\tau)}$ and $f(\bar{u}_k; u_{\star k}^{(\tau)}) = f(u_{\star k}^{(\tau)}; a_k^{(\tau)}), \forall k$.

Because $\bar{u}_k < u_{\star k}^{(\tau)}, \forall k$, the strictly inequality holds in constraint (18a). In addition, since all elements of A_k are positive, there exists a vector $\mathbf{p}'$ such as $|\mathbf{p}'| < |\mathbf{p}^{\star(\tau)}|$ satisfying the first constraint. Now consider a candidate $\left(\mathbf{p}', \mathbf{q}_\star^{(\tau)}, \bar{\mathbf{u}}, \mathbf{v}_\star^{(\tau)}\right)$. This set satisfies all the constraints of $\boldsymbol{\mathcal{P}}_\mathbf{2}(\mathbf{u}_\star^{(\tau)}, \mathbf{b}^{(\tau+1)})$, and therefore is a feasible solution of the optimization problem. Thus, the optimal objective function at the iteration $\tau + 1$ is $P^{(\tau+1)} \leq \text{sum}(\mathbf{p}' + \mathbf{q}_\star^{(\tau)}) < P^{(\tau)}$. Similar conclusion is observed when $v_{\star k}^{(\tau)} < b_k^{(\tau)}$, which completes the proof of the proposition. ∎

Proposition 1 guarantees the convergence of the proposed iterative algorithm in Table I. Although not proving the optimality of the resorted problem (17), Proposition 1 provides the guidance for using the iterative algorithm.

$\boldsymbol{\mathcal{P}}_\mathbf{2}(\mathbf{a}, \mathbf{b})$:

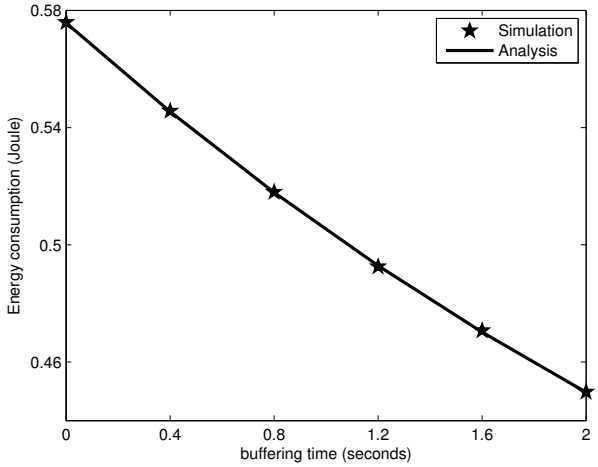

Fig. 2: Energy-buffering time trade-off for single-user case.

$$\underset{\mathbf{p},\mathbf{q},u_k,v_k}{\texttt{Minimize}} \quad \boldsymbol{\beta}^T\mathbf{p} + \boldsymbol{\beta}^T\mathbf{q} \tag{18}$$

$$\texttt{s.t.} \quad \frac{\theta}{T_k}\log_2(A_k\mathbf{p}) + \log_2(A_k\mathbf{q}) \geq r_k + \frac{\theta}{T_k}u_k + v_k, \forall k \in \mathcal{K}', \tag{18a}$$

$$B_k\mathbf{p} \leq e^{a_k}(u_k + 1 - a_k), \forall k \in \mathcal{K}', \tag{18b}$$

$$B_k\mathbf{q} \leq e^{b_k}(v_k + 1 - b_k), \forall k \in \mathcal{K}', \tag{18c}$$

$$\boldsymbol{\beta}^T\mathbf{p} \leq P_{tot} + \sigma^2; \boldsymbol{\beta}^T\mathbf{q} \leq P_{tot} + \sigma^2.$$

V. NUMERICAL RESULTS

This section presents numerical results to demonstrate the derived optimization. The system parameters for simulations are as follows: $\mathcal{B} = 1$ MHz, $\kappa = -20$ dB, $\sigma^2 = -10$ dBm, $Q = 48$ Mbits, and the request rate $r_1 = \cdots = r_K = r = 4$ Mbps which is corresponding to the expected serving time $T = Q/r = 12$ seconds, $P_{tot} = 2$ Watt.

Fig. 2 presents the EBT for the single-user scenario without caching, i.e., $M = 0$. It is observed that the analysis perfectly matches simulation results. If the user does not allow any delay, it costs 0.58 Joule to send the requested file. However, if the user can tolerate a delay of 0.8 seconds, the system can save 10% of the energy cost.

Fig. 3 plots the energy consumption in multi-user systems under two precoding designs for two cases: without caching, i.e., $M = 0$ (left subfigure), and with a cache size $M = 0.1N$ (right subfigure). The energy consumption is calculated based on the optimial solution of the formulated problems in Section IV. It is shown that the MMSE-based design is more efficient than the ZF-based design in the considered setting. In particular, the MMSE design consumes approximately 10% less than the ZF design. It is also shown that with a cache size equal to 10% of the library size, the system can significantly reduce 75% the total energy usage. In all cases, increasing the tolerated latency results in less energy consumption. We would remark that the average energy cost per user in this case (left subfigure) is higher than in the single-user scenario since

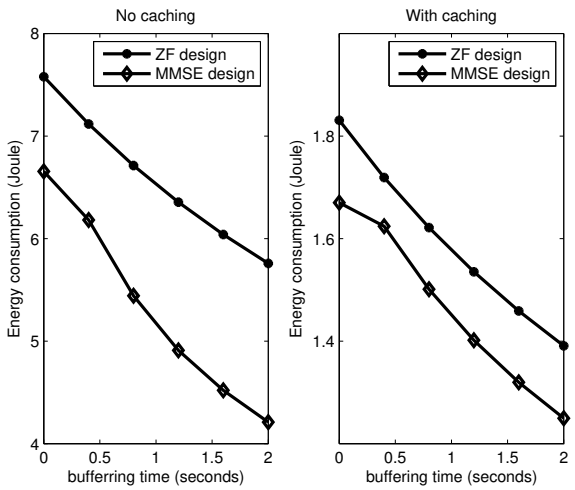

Fig. 3: Total energy consumptions as a function of buffering time for multiple-user cases, $K = 4, L = 5$. Left figure - No caching, i.e., $M = 0$. Right figure - Cache size is 10% the library size, i.e., $M = 0.1N$.

additional energy is required to mitigate inter-user interference.

VI. Conclusions

We have analysed the energy performance of cache-assisted content delivery networks in which a date centre is serving users via shared wireless channels. First, we have derived an energy-buffering time trade-off in a closed-form expression for single-user scenarios. We then have formulated two optimization problems corresponding two linear precoding design for multi-user systems to minimize the total system energy consumption taking into account a allowable latency. The developed framework can be utilized as a guideline for system design and optimization for latency-tolerant services.

Acknowledgement

This work is supported in part by the ERC AGNOSTIC project (R-AGR-3283), the Luxembourg FNR CORE ProCAST project, and the FNR CORE ROSETTA project (11632107).

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

[3] 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.

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

[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] 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.

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

[8] 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.

[9] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Energy-efficient design for edge-caching wireless networks: When is coded-caching beneficial?" in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.*, Jul. 2017, pp. 1–5.

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

[11] 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.

[12] Y. Xu, Y. Li, Z. Wang, T. Lin, G. Zhang, and S. Ci, "Coordinated caching model for minimizing energy consumption in radio access network," in *Proc. IEEE Int. Conf. Commun.*, 2014, pp. 2406–2411.

[13] T. X. Vu, S. Chatzinotas, and B. Ottersten "Edge-Caching Wireless Networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, to appear.

[14] T. X. Vu, S. Chatzinotas, and B. Ottersten "Energy Minimization for Cache-assisted Content Delivery Networks with Wireless Backhaul," *IEEE Wireless Commun. Lett.*, vol. pp, no. pp, pp. 1–1, 2018.

[15] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *IEEE INFOCOM*, Mar. 1999, vol. 1, pp. 126–134.