Decentralized Caching and Coded Delivery over Gaussian Broadcast Channels

Mohammad Mohammadi Amiri and Deniz Gündüz
Electrical and Electronic Engineering Department, Imperial College London, London SW7 2BT, U.K.
Email: {m.mohammadi-amiri15, d.gunduz}@imperial.ac.uk

***Abstract*—A cache-aided K-user Gaussian broadcast channel (BC) is considered. The transmitter has a library of N equal-rate files, from which each user demands one. The impact of the equal-capacity receiver cache memories on the minimum required transmit power to satisfy all user demands is studied. Decentralized caching with uniformly random demands is considered, and both the minimum average power (averaged over all demand combinations) and the minimum peak power (minimum power required to satisfy the worst-case demand combination) are studied. Upper and lower bounds are presented on the minimum required average and peak transmit power as a function of the cache capacity, assuming uncoded cache placement. The gaps between the upper and lower bounds on both the minimum peak and average power values are shown to be relatively small through numerical results, particularly for large cache capacities.**

I. Introduction

Proactive content caching is a promising technique to alleviate growing peak traffic rates in wireless networks by shifting a portion of the peak traffic to off-peak hours. By placing popular contents in local cache memories during off-peak periods, it is possible to reduce the data that each user needs to receive during peak traffic periods [1]. This reduction in rate due to partial availability of contents is called *local caching gain*. It has been shown in [2], [3] that, if user requests are satisfied simultaneously through multicast transmissions, coded delivery can be performed to further reduce the rate of information that needs to be transmitted over the shared communication channel, called the *global caching gain*. Many recent works provide caching and delivery techniques that further improve the global caching gain [4]–[10].

Here we assume that the delivery phase takes place over a noisy channel; in particular we consider a Gaussian broadcast channel (BC) from the server to the users. Several recent papers have studied delivery over noisy channels. Fading and interference channel models are considered in [11] and [12], respectively. In [13] and [14], centralized caching is considered while the delivery phase takes place over a packet-erasure BC. The capacity-memory trade-off is investigated in this setting assuming that only the weak users have caches. Caching over a Gaussian BC is studied in [15] in the high power regime.

Assuming equal-rate files, here we study cache-aided data delivery over a Gaussian BC. Different from [13] and [14], which focus on maximizing the common rate of the contents that can be simultaneously delivered, our goal here is to study the benefits of caching in reducing the transmit power. We assume noiseless cache *placement phase* taking place in a decentralized manner, where a random subset of each file is cached by each user independently. When the user demands are revealed, the *delivery phase* is performed over a Gaussian BC. We are interested in the transmission power required in the delivery phase. Assuming that the files are equally likely to be requested by each user, we first consider the minimum required *average power* to serve all the users, averaged over all possible demand combinations. We then consider the lowest power value required to satisfy the worst-case demand combination, called the *peak power*. We provide upper and lower bounds on the minimum average and peak power values as a function of the file rate and the cache capacity. The lower bound is provided assuming uncoded cache placement phase. The proposed delivery strategy uses superposition coding and power allocation for content delivery.

This research was supported in part by the European Research Council (ERC) through Starting Grant BEACON (agreement #677854).

II. System Model and Preliminaries

We study cache-aided content delivery over a K-user Gaussian BC. The transmitter has a library of N files, $\mathbf{W} \triangleq W_1, ..., W_N$, each distributed uniformly over the set[1] $\left[\left\lceil 2^{nR}\right\rceil\right]$, where R is the rate of the files and n denotes the blocklength. Each user has a cache memory of size MR. Data delivery to users takes place in two phases. Users' caches are filled during the initial *placement phase*, which is performed within a period of low traffic, and without knowing the user demands. The caching function for user k is $\phi_k : \left[\left\lceil 2^{nR}\right\rceil\right]^N \rightarrow \left[\left\lfloor 2^{nMR}\right\rfloor\right]$, which maps the library to the cache content U_k of user k, i.e., $U_k = \phi_k(\mathbf{W})$, for $k \in [K]$.

Each user requests a single file from the library, where W_{d_k}, $d_k \in [N]$, denotes the file requested by user $k \in [K]$. User demands are revealed after the placement phase, and we assume that requests are satisfied simultaneously during the *delivery phase*. It is assumed that user demands are independent and uniformly distributed over the library. For a demand vector $\mathbf{d} = (d_1, ..., d_K)$, the users are served by a common message $X^n(\mathbf{W}, \mathbf{d})$, generated by the delivery function $\psi : \left[\left\lceil 2^{nR}\right\rceil\right]^N \times [N]^K \rightarrow \mathbb{R}^n$. The average power of the channel input $X^n(\mathbf{W}, \mathbf{d})$ is evaluated by $P(\mathbf{W}, \mathbf{d}) \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i^2(\mathbf{W}, \mathbf{d})$. We define the average power constraint satisfied for demand

[1]For any positive integer i, $[i]$ denotes the set $\{1, ..., i\}$.

vector $\mathbf{d}$ as $P(\mathbf{d}) \triangleq \max_{W_1,\ldots,W_N} P(\mathbf{W},\mathbf{d})$. In channel use $i \in [n]$, user $k \in [K]$ receives $Y_{k,i}(\mathbf{W},\mathbf{d})$ through a Gaussian channel

$$Y_{k,i}(\mathbf{W},\mathbf{d}) = X_i(\mathbf{W},\mathbf{d}) + Z_{k,i}, \tag{1}$$

where $X_i(\mathbf{W},\mathbf{d})$ is the ith channel input, and $Z_{k,i}$ is the independent zero-mean Gaussian noise at user k with variance σ_k^2. Without loss of generality, we assume that $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_K^2$. User $k \in [K]$ reconstructs $\hat{W}_{d_k}$ using its channel output $Y_k^n(\mathbf{W},\mathbf{d})$, local cache contents U_k, and demand vector $\mathbf{d}$ through the following function $\mu_k : \mathbb{R}^n \times \left[\left\lfloor 2^{nMR}\right\rfloor\right] \times [N]^K \to \left[\left\lceil 2^{nR}\right\rceil\right]$, where $\hat{W}_{d_k} = \mu_k\left(Y_k^n(\mathbf{W},\mathbf{d}), U_k, \mathbf{d}\right)$. The probability of error is defined as $P_e \triangleq \Pr\left\{\bigcup_{\mathbf{d}\in[N]^K}\bigcup_{k=1}^{K}\left\{\hat{W}_{d_k} \neq W_{d_k}\right\}\right\}$.

An (n,R,M) code for the above caching system consists of K caching functions $\phi_1,\ldots,\phi_K$, channel encoding function ψ, and K decoding functions $\mu_1,\ldots,\mu_K$. We say that an $\left(R,M,\bar{P},\hat{P}\right)$ tuple is achievable if for every $\varepsilon > 0$, there exists an (n,R,M) code which, for n large enough, satisfies $P_e < \varepsilon$, and $\mathrm{E}_{\mathbf{d}}[P(\mathbf{d})] \leq \bar{P}$ and $P(\mathbf{d}) \leq \hat{P}$, $\forall \mathbf{d}$. For given rate R and normalized cache capacity M, the average and peak power-memory trade-offs are defined, respectively, as

$$\bar{P}^*(R,M) \triangleq \inf\left\{\bar{P} : \left(R,M,\bar{P},\infty\right) \text{ is achievable}\right\}, \tag{2a}$$

$$\hat{P}^*(R,M) \triangleq \inf\left\{\hat{P} : \left(R,M,\hat{P},\hat{P}\right) \text{ is achievable}\right\}. \tag{2b}$$

Note that $\bar{P}^*$ is evaluated by allowing a different transmission power for each demand combination, and minimising the average power value across demand combinations. While, $\hat{P}^*$ characterizes the worst-case transmit power, that is, the power that is sufficient to satisfy all possible demand combinations.

The following proposition states the minimum total power required to achieve a given rate tuple in a Gaussian BC without receiver caches, and the corresponding power allocation at the transmitter. This well-known result from multi-user information theory over Gaussian BCs will be instrumental in characterizing our lower and upper bounds.

Proposition 1. *Consider the K-user Gaussian BC presented above with $M = 0$, i.e., no receiver caches. We assume that a distinct message of rate R_k is targeted for user k, for $k \in [K]$. The minimum total power P is achieved by superposition coding with power $\alpha_k P$ allocated for the message of user k, for $k = 1,...,K$, given by [16]*

$$\alpha_k P = \left(2^{2R_k} - 1\right)\left(\sigma_k^2 + \sum_{i=k+1}^{K}\left(\sigma_i^2\left(2^{2R_i} - 1\right)\prod_{j=k+1}^{i-1} 2^{2R_j}\right)\right), \tag{3}$$

and the total transmitted power is

$$P = \sum_{k=1}^{K}\left(\sigma_k^2\left(2^{2R_k} - 1\right)\prod_{i=1}^{k-1} 2^{2R_i}\right). \tag{4}$$

For a demand vector $\mathbf{d}$ in the delivery phase, we denote the number of distinct demands by $N'_{\mathbf{d}}$, where $N'_{\mathbf{d}} \leq \min\{N,K\}$. Let $\mathcal{U}_{\mathbf{d}}$ denote the set of users with distinct requests, which have the worst channel qualities; that is, $\mathcal{U}_{\mathbf{d}}$ consists of $N'_{\mathbf{d}}$ indices corresponding to users with distinct requests, where a user is included in set $\mathcal{U}_{\mathbf{d}}$ iff it has the worst channel quality among all the users with the same demand, i.e.,

$$k \in \mathcal{U}_{\mathbf{d}}, \text{ iff } \sigma_k^2 \geq \sigma_{k'}^2, \ \forall k' \text{ s.t. } d_{k'} = d_k. \tag{5}$$

Note that, for any demand vector $\mathbf{d}$, $1 \in \mathcal{U}_{\mathbf{d}}$. For each user $k \in [K]$, let $\mathcal{U}_{\mathbf{d},k}$ denote the set of users in $\mathcal{U}_{\mathbf{d}}$ which have better channels than user k:

$$\mathcal{U}_{\mathbf{d},k} \triangleq \{i \in \mathcal{U}_{\mathbf{d}} : i > k\}, \quad \text{for } k \in [K]. \tag{6}$$

We denote the cardinality of $\mathcal{U}_{\mathbf{d},k}$ by $N'_{\mathbf{d},k}$, i.e., $N'_{\mathbf{d},k} = |\mathcal{U}_{\mathbf{d},k}|$.

III. Main Results

In this section upper and lower bounds are presented on the optimal average and peak power-memory trade-offs. While the upper bounds on $\bar{P}^*(R,M)$ and $\hat{P}^*(R,M)$ are presented in the following theorem, the achievable scheme that is used to obtain these bounds will be described in Section IV.

Theorem 1. *For decentralized caching in the cache-aided Gaussian BC presented in Section II, we have*

$$\bar{P}^*(R,M) \leq \bar{P}_{\mathrm{UB}}(R,M) \triangleq \mathrm{E}_{\mathbf{d}}\left[\sum_{i=1}^{K}\left(\sigma_i^2\left(2^{2R_{\mathbf{d},i}} - 1\right)\prod_{j=1}^{i-1} 2^{2R_{\mathbf{d},j}}\right)\right], \tag{7a}$$

where, for $k = 1,...,K$,

$$R_{\mathbf{d},k} \triangleq \begin{cases}\left(1-\frac{M}{N}\right)^k R, & \text{if } k \in \mathcal{U}_{\mathbf{d}},\\ \left(1-\frac{M}{N}\right)^k\left(1-\left(1-\frac{M}{N}\right)^{N'_{\mathbf{d},k}}\right)R, & \text{otherwise},\end{cases} \tag{7b}$$

and

$$\hat{P}^*(R,M) \leq \hat{P}_{\mathrm{UB}}(R,M) \triangleq \sum_{i=1}^{\min\{N,K\}}\sigma_i^2\left(2^{2R\left(1-\frac{M}{N}\right)^i} - 1\right)2^{2R\left(\frac{N}{M}-1\right)\left(1-\left(1-\frac{M}{N}\right)^{i-1}\right)}. \tag{8}$$

The average and peak power-memory trade-offs $\bar{P}^*(R,M)$ and $\hat{P}^*(R,M)$, respectively, are lower bounded assuming an uncoded cache placement phase in the following theorem, whose proof can be found in [17].

Theorem 2. *In decentralized caching over the cache-aided Gaussian BC model described in Section II, we have*

$$\bar{P}^*(R,M) \geq \bar{P}_{\mathrm{LB}}(R,M) \triangleq \mathrm{E}_{\mathcal{U}_{\mathbf{d}}}\left[\sum_{i=1}^{N'_{\mathbf{d}}}\left(\sigma^2_{\pi_{\mathcal{U}_{\mathbf{d}}}(i)}\left(2^{2R\left(1-\min\left\{MN'_{\mathbf{d}}/N,1\right\}\right)} - 1\right)2^{2(i-1)R\left(1-\min\left\{MN'_{\mathbf{d}}/N,1\right\}\right)}\right)\right], \tag{9}$$

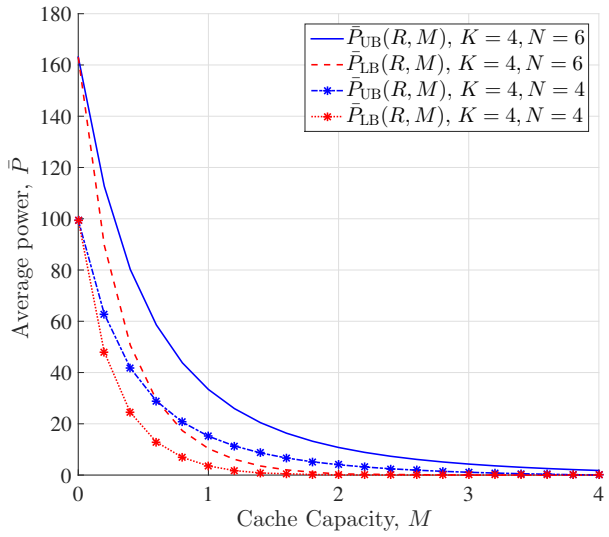

Fig. 1. Average power-memory trade-off for a Gaussian BC with $K=4$ users, and $N=4$ and $N=6$ files in the library. Noise variance at user k is $\sigma_k^2=2-0.2(k-1)$, for $k=1,...,4$, and the file rate is fixed to $R=1$.

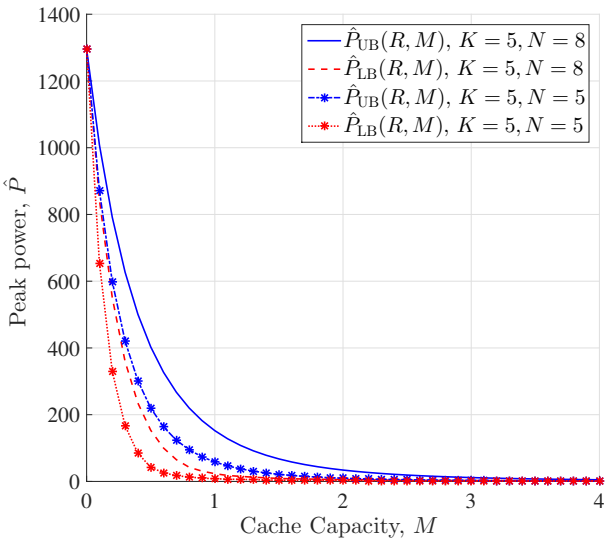

Fig. 2. Peak power-memory trade-off for a Gaussian BC with $K=5$ users, and $N=5$ and $N=8$ files in the library. Noise variance at user k is $\sigma_k^2=2-0.2(k-1)$, for $k=1,...,5$, and the file rate is fixed to $R=1$.

where $\mathrm{E}_{\mathcal{U}_{\mathbf{d}}}[\cdot]$ *takes the expectation over all possible sets* $\mathcal{U}_{\mathbf{d}}$*, and* $\pi_{\mathcal{S}}$ *is a permutation over set* $\mathcal{S}\subset[K]$*, such that* $\sigma^2_{\pi_{\mathcal{S}}(1)}\geq\sigma^2_{\pi_{\mathcal{S}}(2)}\geq\cdots\geq\sigma^2_{\pi_{\mathcal{S}}(|\mathcal{S}|)}$*. We have the following lower bound on the minimal required peak transmit power:*

$$\hat{P}^*(R,M)\geq\hat{P}_{\rm LB}(R,M)\triangleq \max_{\mathcal{S}\subset[\min\{N,K\}]}\left\{\sum_{i=1}^{|\mathcal{S}|}\left(\sigma^2_{\pi_{\mathcal{S}}(i)}\left(2^{2R(1-\min\{M|\mathcal{S}|/N,1\})}-1\right)2^{2(i-1)R(1-\min\{M|\mathcal{S}|/N,1\})}\right)\right\}. \quad (10)$$

In Fig. 1, upper and lower bounds on the average power-memory trade-off $\bar{P}^*(R,M)$, presented in (7) and (9), respectively, are plotted, for $K=4$ users, and considering $N=4$ and $N=6$ files in the library. The rate of the files in the library is fixed to $R=1$, and the noise variance at user k is $\sigma_k^2=2-0.2(k-1)$, for $k=1,\ldots,4$. We observe that the minimum average power drops very quickly even with a small cache capacity available at the users. The upper and lower bounds meet for the trivial case of zero cache capacity, but there is a gap between the two for small M. The gap diminishes as the cache capacity increases, which shows that the proposed coded delivery scheme is near optimal for high cache capacities. We also observe that the gap between the bounds increases with the number of files in the library.

In Fig. 2, upper and lower bounds on the peak power-memory trade-off $\hat{P}^*(R,M)$, presented in (8) and (10), respectively, are plotted. The file rate is again $R=1$. The number of users is $K=5$, and we consider $N=5$ and $N=8$ files. The noise variance at user k is $\sigma_k^2=2-0.2(k-1)$, for $k=1,\ldots,5$. The peak power values exhibit similar behaviour to the average power, with significantly higher values. This shows that adapting the transmit power to the demand combination can save a significant amount of energy.

IV. Proposed Caching and Delivery Scheme

Here we present a cache-aided coded delivery scheme, which uses superposition transmission of coded packets together with power allocation, and achieves the average and peak power-memory trade-offs in (7) and (8), respectively. For simplicity, we assume that both nR and nMR are integers.

A. Placement phase

Decentralized uncoded cache placement is performed [3], where each user caches nMR/N random bits of each file of length nR bits independently. Since there are a total of N files in the library, the cache capacity constraint is satisfied. The part of file i cached exclusively by the users in set $\mathcal{S}\subset[K]$ is denoted by $W_{i,\mathcal{S}}$, for $i=1,...,N$. For n large enough, the rate of $W_{i,\mathcal{S}}$ can be approximated by

$$\left(\frac{M}{N}\right)^{|\mathcal{S}|}\left(1-\frac{M}{N}\right)^{K-|\mathcal{S}|}R. \quad (11)$$

The cache content at user k is given by

$$U_k=\bigcup_{i\in[N]}\bigcup_{\mathcal{S}\subset[K]:k\in\mathcal{S}}W_{i,\mathcal{S}}. \quad (12)$$

B. Delivery phase

Consider any non-empty set of users $\mathcal{S}\subset[K]$. For a demand vector $\mathbf{d}$, by delivering the coded message

$$V_{\mathcal{S}}\triangleq\bigoplus_{k\in\mathcal{S}}W_{d_k,\mathcal{S}\setminus\{k\}} \quad (13)$$

of rate

$$\left(\frac{M}{N}\right)^{|\mathcal{S}|-1}\left(1-\frac{M}{N}\right)^{K-|\mathcal{S}|+1}R \quad (14)$$

to users in $\mathcal{S}$, each user $p \in \mathcal{S}$ can recover subfile $W_{d_p,\mathcal{S}\backslash\{p\}}$, since it knows all the subfiles $W_{d_q,\mathcal{S}\backslash\{q\}}$, $\forall q \in \mathcal{S}\backslash\{p\}$, where $\oplus$ represents the bitwise XOR operation. The demand of each user $k \in [K]$ can be satisfied after receiving $\bigcup_{\mathcal{S}:k\in\mathcal{S}} V_{\mathcal{S}}$ together with its cache content. Thus, having delivered all coded messages $\bigcup_{\mathcal{S}\subset[K]} V_{\mathcal{S}}$ to the users in $\mathcal{S}$, demand vector $\mathbf{d}$ can be satisfied. However, as proposed in [9], for a demand vector $\mathbf{d}$ with $N'_{\mathbf{d}}$ distinct requests, if $N'_{\mathbf{d}} < K$, there is at least one file requested by more than one user, and not all coded messages $V_{\mathcal{S}}$, $\forall \mathcal{S} \subset [K]$, are needed to be delivered.

Following [9, Lemma 1], for a demand vector $\mathbf{d}$ with $N'_{\mathbf{d}}$ distinct requests and set $\mathcal{U}_{\mathbf{d}}$, let $\mathcal{B}$ be a subset of $[K]$ satisfying $\mathcal{U}_{\mathbf{d}} \subset \mathcal{B}$. We define $\mathcal{G}_{\mathcal{B}}$ as the set consisting of all subsets of $\mathcal{B}$ with cardinality $N'_{\mathbf{d}}$, such that all $N'_{\mathbf{d}}$ users in each subset request distinct files. For any set $\mathcal{B}$, we have

$$\bigoplus\nolimits_{\mathcal{G}\in\mathcal{G}_{\mathcal{B}}} V_{\mathcal{B}\backslash\mathcal{G}} = \mathbf{0}, \tag{15}$$

where $\mathbf{0}$ denotes the all-zero vector.

Remark 1. *Given a demand vector $\mathbf{d}$ with set $\mathcal{U}_{\mathbf{d}}$ of size $N'_{\mathbf{d}} < K$, and for any set $\mathcal{S} \subset [K]\backslash\mathcal{U}_{\mathbf{d}}$ of users, by setting $\mathcal{B} = \mathcal{S}\cup\mathcal{U}_{\mathbf{d}}$, we have*

$$\begin{aligned}\bigoplus\nolimits_{\mathcal{G}\in\mathcal{G}_{\mathcal{B}}} V_{\mathcal{B}\backslash\mathcal{G}} &= \left(\bigoplus\nolimits_{\mathcal{G}\in\mathcal{G}_{\mathcal{B}}\backslash\mathcal{U}_{\mathbf{d}}} V_{\mathcal{B}\backslash\mathcal{G}}\right) \oplus V_{\mathcal{B}\backslash\mathcal{U}_{\mathbf{d}}} \\ &= \left(\bigoplus\nolimits_{\mathcal{G}\in\mathcal{G}_{\mathcal{B}}\backslash\mathcal{U}_{\mathbf{d}}} V_{\mathcal{B}\backslash\mathcal{G}}\right) \oplus V_{\mathcal{S}} = \mathbf{0}, \end{aligned} \tag{16}$$

which leads to

$$V_{\mathcal{S}} = \bigoplus\nolimits_{\mathcal{G}\in\mathcal{G}_{\mathcal{B}}\backslash\mathcal{U}_{\mathbf{d}}} V_{\mathcal{B}\backslash\mathcal{G}}. \tag{17}$$

Thus, having received all the coded messages $V_{\mathcal{B}\backslash\mathcal{G}}$, $\forall \mathcal{G} \in \mathcal{G}_{\mathcal{B}}\backslash\mathcal{U}_{\mathbf{d}}$, $V_{\mathcal{S}}$ can be recovered through (17)*. Note that, for any $\mathcal{G} \in \mathcal{G}_{\mathcal{B}}\backslash\mathcal{U}_{\mathbf{d}}$, we have*

$$|\mathcal{B}\backslash\mathcal{G}| = |\mathcal{S}|, \tag{18a}$$

$$(\mathcal{B}\backslash\mathcal{G}) \cap \mathcal{U}_{\mathbf{d}} \neq \emptyset, \tag{18b}$$

that is, each coded message on the right hand side of (17) *is targeted for a set of $|\mathcal{S}|$ users, at least one of which is in set $\mathcal{U}_{\mathbf{d}}$. Furthermore, for each $k \in \mathcal{S}$, there is a user $k' \in \mathcal{U}_{\mathbf{d}}$ with $\sigma^2_{k'} \geq \sigma^2_k$, such that $d_{k'} = d_k$. Note that no two users with the same demand are in any set of $\mathcal{G} \in \mathcal{G}_{\mathcal{B}}$. Thus, for any $\mathcal{G} \in \mathcal{G}_{\mathcal{B}}\backslash\mathcal{U}_{\mathbf{d}}$, either $k \in \mathcal{B}\backslash\mathcal{G}$ or $k' \in \mathcal{B}\backslash\mathcal{G}$.*

Given a demand vector $\mathbf{d}$, the delivery phase is designed such that only the coded messages $V_{\mathcal{S}}$, $\forall \mathcal{S} \subset [K]$ that satisfy $\mathcal{S} \cap \mathcal{U}_{\mathbf{d}} \neq \emptyset$, are delivered, i.e., the coded messages that are targeted for at least one user in $\mathcal{U}_{\mathbf{d}}$ are delivered, and the remaining coded messages can be recovered through (17). To achieve this, for any such set $\mathcal{S}$ with $\mathcal{S} \cap \mathcal{U}_{\mathbf{d}} \neq \emptyset$, the transmission power is adjusted such that the worst user in $\mathcal{S}$ can decode it; and so can all the other users in $\mathcal{S}$ due to the degradedness of the Gaussian BC. Therefore, the demand of every user in $\mathcal{U}_{\mathbf{d}}$ is satisfied.

The main technique to deliver the coded messages is to start from the worst user, i.e., user $1 \in \mathcal{U}_{\mathbf{d}}$, and transmit the coded messages targeted for it. When the worst user receives all its intended coded messages, we target the next worst user, and send it the coded messages targeted for it, which have not been already delivered taking into account the fact that only those coded messages $V_{\mathcal{S}}$ with $\mathcal{S} \cap \mathcal{U}_{\mathbf{d}} \neq \emptyset$ are delivered, and so on so forth.

For a demand vector $\mathbf{d}$, all the coded messages $\bigcup_{\mathcal{S}\subset[K]:1\in\mathcal{S}} V_{\mathcal{S}}$ are delivered such that user 1 can decode them. Thus, the total rate targeted to user 1 is given by

$$\begin{aligned} R_{\mathbf{d},1} &= \sum_{i=0}^{K-1} \binom{K-1}{i} \left(\frac{M}{N}\right)^i \left(1-\frac{M}{N}\right)^{K-i} R \\ &= \left(1-\frac{M}{N}\right) R. \end{aligned} \tag{19}$$

The transmission power of the message for user 1 is adjusted such that it can decode $\bigcup_{\mathcal{S}\subset[K]:1\in\mathcal{S}} V_{\mathcal{S}}$. Accordingly, all the other users can decode it, and obtain the bits of their requests placed in $\bigcup_{\mathcal{S}\subset[K]:1\in\mathcal{S}} V_{\mathcal{S}}$ in XOR-ed form. A similar procedure is performed for users 2 to K. For each user $k \in [2:K]$, if $k \in \mathcal{U}_{\mathbf{d}}$, coded contents

$$\bigcup_{\mathcal{S}\subset[k:K]:k\in\mathcal{S}} V_{\mathcal{S}} \tag{20}$$

are delivered. On the other hand, if $k \notin \mathcal{U}_{\mathbf{d}}$, coded contents

$$\bigcup_{\mathcal{S}\subset[k:K]:\mathcal{S}\cap\mathcal{U}_{\mathbf{d}}\neq\emptyset,k\in\mathcal{S}} V_{\mathcal{S}}, \tag{21}$$

which are equivalent to

$$\bigcup_{\mathcal{S}\subset[k:K]:k\in\mathcal{S}} V_{\mathcal{S}} - \bigcup_{\mathcal{S}\subset[k:K]\backslash\mathcal{U}_{\mathbf{d},k},k\in\mathcal{S}} V_{\mathcal{S}}, \tag{22}$$

are delivered. Thus, if $k \in \mathcal{U}_{\mathbf{d}}$ the total rate targeted to user $k \in [2:K]$ is

$$\begin{aligned} R_{\mathbf{d},k} &= \sum_{i=0}^{K-k} \binom{K-k}{i} \left(\frac{M}{N}\right)^i \left(1-\frac{M}{N}\right)^{K-i} R \\ &= \left(1-\frac{M}{N}\right)^k R. \end{aligned} \tag{23}$$

However, if $k \notin \mathcal{U}_{\mathbf{d}}$ the total rate targeted to user $k \in [2:K]$ is given by

$$\begin{aligned} R_{\mathbf{d},k} &= \sum_{i=0}^{K-k} \binom{K-k}{i} \left(\frac{M}{N}\right)^i \left(1-\frac{M}{N}\right)^{K-i} R \\ &\quad - \sum_{i=0}^{K-k-N'_{\mathbf{d},k}} \binom{K-k-N'_{\mathbf{d},k}}{i} \left(\frac{M}{N}\right)^i \left(1-\frac{M}{N}\right)^{K-i} R \\ &= \left(1-\frac{M}{N}\right)^k \left(1-\left(1-\frac{M}{N}\right)^{N'_{\mathbf{d},k}}\right) R. \end{aligned} \tag{24}$$

In total, to deliver coded message $V_{\mathcal{S}}$ to the worst user in $\mathcal{S}$, for any non-empty set $\mathcal{S} \subset [K]$, such that $\mathcal{S} \cap \mathcal{U}_{\mathbf{d}} \neq \emptyset$, a total

rate of

$$R_{\mathbf{d},k} = \begin{cases} \left(1-\frac{M}{N}\right)^k R, & \text{if } k \in \mathcal{U}_{\mathbf{d}}, \\ \left(1-\frac{M}{N}\right)^k \left(1-\left(1-\frac{M}{N}\right)^{N'_{\mathbf{d},k}}\right) R, & \text{otherwise}, \end{cases} \tag{25}$$

is targeted to user k. In this case, each user $k \in \mathcal{S}$ can obtain $V_{\mathcal{S}}$, $\forall \mathcal{S} \subset [K]$ that satisfy $\mathcal{S} \cap \mathcal{U}_{\mathbf{d}} \neq \emptyset$. Thus, the demands of users in $\mathcal{U}_{\mathbf{d}}$ are satisfied.

Next, we illustrate that the users in $[K] \backslash \mathcal{U}_{\mathbf{d}}$ can decode their requested files without being delivered any extra messages. Given any set of users $\mathcal{S}$ such that $\mathcal{S} \cap \mathcal{U}_{\mathbf{d}} = \emptyset$, we need to show that every user in $\mathcal{S}$ can decode all coded messages $V_{\mathcal{B}\backslash\mathcal{G}}$, $\forall \mathcal{G} \in \mathcal{G}_{\mathcal{B}} \backslash \mathcal{U}_{\mathbf{d}}$, where $\mathcal{B} = \mathcal{S} \cup \mathcal{U}_{\mathbf{d}}$. According to (18b), there is at least one user in $\mathcal{U}_{\mathbf{d}}$ in any set of users $\mathcal{B}\backslash\mathcal{G}$, $\forall \mathcal{G} \in \mathcal{G}_{\mathcal{B}} \backslash \mathcal{U}_{\mathbf{d}}$. Thus, all coded messages $V_{\mathcal{B}\backslash\mathcal{G}}$ have been delivered. Remember the fact that, for each user $k \in \mathcal{S}$, either $k \in \mathcal{B}\backslash\mathcal{G}$ or $k' \in \mathcal{B}\backslash\mathcal{G}$, where $d_{k'} = d_k$ and $k' \in \mathcal{U}_{\mathbf{d}}$, i.e., user k' has a worse channel quality than k, $\forall \mathcal{G} \in \mathcal{G}_{\mathcal{B}} \backslash \mathcal{U}_{\mathbf{d}}$. If $k \in \mathcal{B}\backslash\mathcal{G}$, user k can obtain $V_{\mathcal{B}\backslash\mathcal{G}}$, and if $k' \in \mathcal{B}\backslash\mathcal{G}$, user k' with a worse link can decode $V_{\mathcal{B}\backslash\mathcal{G}}$, which concludes that user k can also decode it due to the degradedness of the Gaussian BC. Thus, each user $k \in \mathcal{S}$ can decode $V_{\mathcal{S}}$ successfully, $\forall \mathcal{S} \subset [K]$ that satisfy $\mathcal{S} \cap \mathcal{U}_{\mathbf{d}} = \emptyset$. This fact illustrates that the demand of the users in $[K] \backslash \mathcal{U}_{\mathbf{d}}$ can also be satisfied.

For any demand vector $\mathbf{d}$ in the delivery phase, we need to deliver a message of rate $R_{\mathbf{d},k}$, given in (25), to user k, for $k = 1, ..., K$. From Proposition 1, the corresponding minimum required power is found to be

$$P_{\mathrm{UB}}\left(R, M, \mathbf{d}\right) \triangleq \sum_{i=1}^{K} \left(\sigma_i^2 \left(2^{2R_{\mathbf{d},i}} - 1\right) \prod_{j=1}^{i-1} 2^{2R_{\mathbf{d},j}} \right). \tag{26}$$

Thus, the average power-memory trade-off for the proposed coded delivery scheme with superposition coding is given by $\mathrm{E}_{\mathbf{d}}\left[P_{\mathrm{UB}}\left(R, M, \mathbf{d}\right)\right] = \bar{P}_{\mathrm{UB}}(R, M)$ stated in Theorem 1.

With the proposed caching scheme, the following peak power performance can be achieved

$$\hat{P}_{\mathrm{UB}}\left(R, M\right) = \max_{\mathbf{d}} \left\{ P_{\mathrm{UB}}\left(R, M, \mathbf{d}\right) \right\}. \tag{27}$$

Observe that, for demand vectors with the same set of users $\mathcal{U}_{\mathbf{d}}$, the required power $P_{\mathrm{UB}}\left(R, M, \mathbf{d}\right)$ is the same. Let $\mathcal{D}_{\mathcal{U}_{\mathbf{d}}}$ be the set of all demand vectors with the same set of users $\mathcal{U}_{\mathbf{d}}$. We define $P_{\mathrm{UB}}\left(R, M, \mathcal{D}_{\mathcal{U}_{\mathbf{d}}}\right)$ as the required power $P_{\mathrm{UB}}\left(R, M, \mathbf{d}\right)$ for any demand vector $\mathbf{d} \in \mathcal{D}_{\mathcal{U}_{\mathbf{d}}}$. From (27), we have

$$\hat{P}_{\mathrm{UB}}\left(R, M\right) = \max_{\mathcal{U}_{\mathbf{d}}} \left\{ P_{\mathrm{UB}}\left(R, M, \mathcal{D}_{\mathcal{U}_{\mathbf{d}}}\right) \right\}. \tag{28}$$

It is shown in [17] that the worst-case demand combination happens when the first $\min\{N, K\}$ users request distinct files, in which case, we have

$$\mathcal{U}_{\mathbf{d}} = [\min\{N, K\}] \tag{29a}$$

$$N'_{\mathbf{d},k} = \begin{cases} \min\{N, K\} - k, & \text{if } k \in \mathcal{U}_{\mathbf{d}}, \\ 0, & \text{otherwise}, \end{cases} \tag{29b}$$

and $\hat{P}_{\mathrm{UB}}\left(R, M\right)$ is found as in (8).

V. Conclusions

We have considered cache-aided content delivery over a Gaussian BC, where the receivers are equipped with equal-size cache memories. We have proposed a decentralized caching and coded delivery scheme, in which the coded contents are transmitted using superposition coding and power allocation. For a given rate for the contents in the library, and imposing the transmitter to satisfy all the user demands reliably, we have studied both the minimum required *peak transmission power*, which is the transmit power constraint that is sufficient to satisfy all user demand combinations, and the minimum *average power* across different user demand combinations, assuming uniform demand distributions. We have provided lower and upper bounds on the required peak and average transmission power values, where the lower bound is derived assuming uncoded cache placement. Our results indicate that, even a small cache capacity at the receivers can provide a significant reduction in the transmission power highlighting the benefits of content caching in increasing the energy efficiency of wireless networks.

References

[1] 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, Aug. 1996.

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

[3] ——, "Decentralized caching attains order optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Apr. 2014.

[4] Z. Chen, P. Fan, and K. B. Letaief, "Fundamental limits of caching: Improved bounds for users with small buffers," *IET Communications*, vol. 10, no. 17, pp. 2315–2318, Nov. 2016.

[5] M. Mohammadi Amiri and D. Gündüz, "Improved delivery rate-cache capacity trade-off for centralized coded caching," in *Proc. IEEE ISITA*, Monterey, CA, Oct.-Nov. 2016, pp. 26–30.

[6] ——, "Fundamental limits of coded caching: Improved delivery rate-cache capacity trade-off," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 806–815, Feb. 2017.

[7] M. Mohammadi Amiri, Q. Yang, and D. Gündüz, "Coded caching for a large number of users," in *IEEE ITW*, Cambridge, UK, Sep. 2016.

[8] K. Wan, D. Tuninetti, and P. Piantanida, "On caching with more users than files," in *IEEE Int'l Sym. Inf. Theory*, Barcelona, Spain, Jul. 2016.

[9] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "The exact rate-memory tradeoff for caching with uncoded prefetching," *arXiv:1609.07817v1 [cs.IT]*, Sep. 2016.

[10] C. Tian, "Symmetry, outer bounds, and code constructions: A computer-aided investigation on the fundamental limits of caching," *arXiv:1611.00024v1 [cs.IT]*, Oct. 2016.

[11] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel," *arXiv:1504.01452v1 [cs.IT]*, Apr. 2015.

[12] N. Naderializadeh, M. Maddah-Ali, and S. Avestimehr, "Fundamental limits of cache-aided interference management," *arXiv:1602.04207v1 [cs.IT]*, Feb. 2016.

[13] S. Saeedi Bidokhti, R. Timo, and M. Wigger, "Noisy broadcast networks with receiver caching," *arXiv:1605.02317v1 [cs.IT]*, May 2016.

[14] M. Mohammadi Amiri and D. Gündüz, "Cache-aided data delivery over erasure broadcast channels," *arXiv:1702.05454v1 [cs.IT]*, Feb. 2017.

[15] J. Zhang and P. Elia, "Wireless coded caching: A topological perspective," *arXiv:1606.08253v1 [cs.IT]*, Jun. 2016.

[16] P. Bergmans, "Coding theorem for broadcast channels with degraded components," *IEEE Trans. Inform. Theory*, vol. 19, no. 2, pp. 197–207, Mar. 1973.

[17] M. Mohammadi Amiri and D. Gündüz, "Caching and coded delivery over Gaussian broadcast channels with power allocation," *in preparation*.