Opportunistic Content Delivery in Fading Broadcast Channels

Asma Ghorbel, Khac-Hoang Ngo, Richard Combes, Mari Kobayashi, Sheng Yang

Opportunistic Content Delivery in Fading Broadcast Channels

A. Ghorbel, K. H. Ngo, R. Combes, M. Kobayashi, and S. Yang
LSS, CentraleSupélec
Gif-sur-Yvette, France
{firstname.lastname}@centralesupelec.fr

***Abstract*—We consider content delivery over fading broadcast channels. A server wants to transmit K files to K users, each equipped with a cache of finite size. Using the coded caching scheme of Maddah-Ali and Niesen, we design an opportunistic delivery scheme where the long-term sum content delivery rate scales with K the number of users in the system. The proposed delivery scheme combines superposition coding together with appropriate power allocation across sub-files intended to different subsets of users. We analyze the long-term average sum content delivery rate achieved by two special cases of our scheme: a) a selection scheme that chooses the subset of users with the largest weighted rate, and b) a baseline scheme that transmits to K users using the scheme of Maddah-Ali and Niesen. We prove that coded caching with appropriate user selection is scalable since it yields a linear increase of the average sum content delivery rate.**

I. Introduction

Content delivery applications such as video streaming are envisionned to represent nearly 75% of the mobile data traffic by 2020. The skewness of the video traffic together with the ever-growing cheap on-board storage memory suggests that the quality of experience can be improved by caching popular content close to the end-users in wireless networks. Recent works have studied the gains provided by caching under various models and assumptions (see e.g. [1], [2] and references therein). In this work, we consider content delivery using coded caching where a server is connected to K users each equipped with a cache of finite memory [1]. A striking result of [1] is that the total number of multicast transmissions to satisfy K distinct requests converges to a constant in the regime of a large K, thus yielding a scalable system.

Substantial effort have been devoted to quantify the gains of coded caching in more realistic scenarios (see e.g. [1, Section VIII], [4]). In particular, some authors have studied coded caching over wireless channels by relaxing the initial assumption of a perfect shared link between the server and users [5]–[9]. It is noted that the performance of coded caching strongly depends on the multicast rate of the underlying wireless channels and the latter is limited by the user in the worst channel condition. Such limitation has been highlighted in [7] which shows that the sum content delivery rate is no longer scalable, if the multicast rate vanishes when $K \to \infty$. This is typically the case for the i.i.d. Rayleigh fading channels (i.i.d. across users and time) [10].

This work was supported by Huawei Technologies France SASU.

In a more realistic scenario where users have asymmetric fading statistics (e.g. in a cellular system), the performance degradation becomes substantial in the sense that most of the resources are allocated to users with low channel quality. To overcome these drawbacks, schemes using multiple antennas [7]–[9] and interference management techniques [5], [6] have been proposed. In this work, we take a different approach based on user scheduling in order to address the following fundamental question that has been overlooked in existing works: *how to exploit the wireless channels opportunistically for content delivery?*

To answer this question, we consider the K-user Gaussian fading broadcast channel with $2^K - 1$ independent messages, each intended to a subset of users, and solve the weighted sum rate maximization in section III. The optimal strategy combines superposition coding with an appropriate power allocation across different messages. The solution at hand can be applied to various communication contexts such as a queued content delivery network [11]. We apply this solution to maximize the sum content delivery rate, assuming that content placement is performed by existing schemes [1], [3]. We analyze the performance of our scheme in two special cases of interest: a) a selection scheme that chooses the subset of users with the largest instantaneous weighted rate, and b) a baseline scheme that applies coded caching to K users. We prove that the selection scheme achieves a linear increase of the average sum content delivery rate in the regime of a large K thus yields a scalable solution. On the other hand, both the baseline and the selection schemes achieve the same sum delivery rate in the high SNR regime, since it is nearly optimal to perform coded caching over all K users in this regime. Moreover, we provide a simple threshold-based feedback scheme which yields the same performance as the selection scheme in the large K regime, while requiring each user to feedback only one bit rather than its channel state information. Numerical examples in Section V show that the linear gain in sum content delivery rate occurs even for relatively small number of users. Proofs of Theorem 1, Propositions 1 and 2 are presented in Appendix.

We use the following notation: $[k] = \{1, \ldots, k\}$, and $f(x) \sim g(x)$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

II. System Model

We consider a content delivery system where a server with N files wishes to transmit K requested files to K users over

a wireless downlink channel. We assume that N files are of equal size of F bits and equal popularity, while each user has a cache of size MF bits, where $M \geq 1$ denotes the cache size measured in files. We define m the normalized cache size denoted by $m = M/N$. Each user can store any part of any file in her cache, by prefetching them during off-peak hours, prior to the actual request, according to centralized or decentralized placement strategies proposed in the literature.

In the decentralized placement of [3], each user independently caches a subset of mF bits of file i, chosen uniformly at random for $i = 1, \ldots, N$ under the memory constraint of MF bits. By letting $W_{i|\mathcal{J}}$ denote the sub-file of W_i stored exclusively in the cache memories of the user set $\mathcal{J}$, the cache memory Z_k of user k after decentralized placement is given by

$$Z_k = \{W_{i\,|\,\mathcal{J}} : \ \forall \mathcal{J} \subseteq [K], \forall \mathcal{J} \ni k, \forall i \in [N]\}. \tag{1}$$

In the centralized cache placement [1], each file is split into $\binom{K}{b}$ disjoint sub-files of equal size, where $b \triangleq \lfloor mK \rfloor$. Each sub-file is cached by users in a subset $\mathcal{J}$ of cardinality $|\mathcal{J}| = b$. The resulting cache memory Z_k is the same as (1) except that the subsets are now restricted to those with a specific cardinality b. Once the requests from users are revealed, the server generates and sequentially conveys the codewords intended to each subset of users. Namely, assuming that user k requests file k for all k, the codeword intended to the subset $\mathcal{J}$ is given by

$$V_{\mathcal{J}} = \oplus_{k \in \mathcal{J}} W_{k|\mathcal{J}\setminus\{k\}}, \tag{2}$$

where $\oplus$ denotes the bit-wise XOR operation. The main idea here is to create a codeword useful to a subset of users by exploiting the receiver side information established during the placement phase.

It has been shown in [1], [3] that the total number of multicast transmissions needed to satisfy K *distinct* demands over the error-free shared link is as follows.

$$T(m,K) = \begin{cases} (1-m)\,\frac{1}{1/K+m}, & \text{centralized caching,} \\ (1-m)\,\frac{1-(1-m)^K}{m}, & \text{decentralized caching.} \end{cases} \tag{3}$$

In the physical layer, we consider the quasi-static Rayleigh fading broadcast channel. The output of user k at channel use t is given by

$$y_k[t] = \sqrt{h_k} x[t] + w_k[t], \tag{4}$$

where x is the input symbol satisfying the power constraint $\frac{1}{n}\sum_{t=1}^{n} |x[t]|^2 \leq P$; $\{h_k\}$ are the fading gains, independently and exponentially distributed $\sim \text{Exp}(1/\gamma_k)$ with mean γ_k; $w_k(t) \sim \mathcal{N}_{\mathbb{C}}(0,1)$ is additive white Gaussian noise assumed independent between users. We assume that $\{h_k\}$ are known by the server and all users.

It is well-known that the multicast capacity of the channel at hand, or the common message rate, is given by

$$R_{\text{mc}}(\boldsymbol{h}) = \log\left(1 + P \min_{j \in [K]} h_j\right) \tag{5}$$

and is limited by the user in the worst fading condition. It has been proved in [7] that such limitation is detrimental for a scalable content delivey network.

To see this, let us first define the sum content delivery rate when coded caching is applied directly to the fading broadcast channel. In order to satisfy the distinct demands from K users, that is to *complete* the transfer of KF demanded bits, one needs to send $T(m,K)F$ bits over the wireless link. The corresponding transmission takes $\frac{T(m,K)F}{R_{\text{mc}}(\boldsymbol{h})}$ units of time. As a result, the sum content delivery rate of a naive application of coded caching for a given channel realization $\boldsymbol{h}$ is given by

$$\frac{K}{T(m,K)} R_{\text{mc}}(\boldsymbol{h})$$

measured in [nats/second/Hz]. We call this scheme the "baseline" scheme, and its long-term average sum content delivery rate is

$$R_{\text{sum,bl}}(K) = \frac{K}{T(m,K)} \mathbb{E}[R_{\text{mc}}(\boldsymbol{h})]. \tag{6}$$

In the case of symmetric fading statistics ($\gamma_k = 1, \forall k$), since the average multicast capacity vanishes as $1/K$ for a large K [10], the average sum content delivery rate converges to a constant, yielding a non-scalable system. This negative result calls for a careful design of content delivery that benefits from the time varying nature of the underlying fading broadcast channel.

III. Problem Formulation

In this section, we study the fading Gaussian broadcast channel where the transmitter wishes to convey $2^K - 1$ mutually independent messages, each intended to a subset of users. We characterize the capacity region of these messages and then solve explicitly the weighted sum rate maximization problem. We show that this formulation allows to maximize the content delivery rate by opportunistically exploiting the wireless channel.

A. Broadcasting private and multiple common messages

We start by observing that the channel at hand in (4) for a given channel realization $\boldsymbol{h}$ corresponds to a stochastically degraded Gaussian broadcast channel. Without loss of generality, let us assume $h_1 \geq \cdots \geq h_K$. The capacity region of the degraded broadcast channel for K private messages and a common message is well-known [12]. In this section, we consider a more general setup where the transmitter wishes to convey $2^K - 1$ mutually independent messages, denoted by $\{M_{\mathcal{J}}\}$, where $M_{\mathcal{J}}$ denotes the message intended to the users in subset $\mathcal{J} \subseteq [K]$. Each user k must decode all messages $\{M_{\mathcal{J}}\}$ for $\mathcal{J} \ni k$. By letting $R_{\mathcal{J}}$ denote the multicast rate of the message $M_{\mathcal{J}}$, we say that the rate-tuple $\boldsymbol{R} \in \mathbb{R}_+^{2^K-1}$ is achievable if there exists encoding and decoding functions which guarantee a rate greater than $\boldsymbol{R}$. The capacity region is defined as the supremum of the achievable rate-tuple. Then we have the following result.

Theorem 1. *The capacity region $\Gamma(\boldsymbol{h})$ of a K-user degraded Gaussian broadcast channel with fading gains $h_1 \geq \cdots \geq h_K$ and $2^K - 1$ independent messages $\{M_{\mathcal{J}}\}$ is given by*

$$R_1 \leq \log(1 + h_1 \alpha_1 P) \tag{7}$$

$$\sum_{\mathcal{K}: k \in \mathcal{K} \subseteq [k]} R_{\mathcal{K}} \leq \log \frac{1 + h_k \sum_{j=1}^{k} \alpha_j P}{1 + h_k \sum_{j=1}^{k-1} \alpha_j P} \quad k = 2, \ldots, K \tag{8}$$

for non-negative variables $\{\alpha_k\}$ such that $\sum_{k=1}^{K} \alpha_k \leq 1$.

Proof: See Appendix VII-A. ∎

The achievability builds on superposition coding at the transmitter and successive interference cancellation at receivers. For $K = 3$, the transmit signal is simply given by

$$x = x_1 + x_2 + x_3 + x_{12} + x_{23} + x_{13} + x_{123},$$

where $\{x_{\mathcal{J}}\}$ are mutually independent and $x_{\mathcal{J}} \sim \mathcal{N}_{\mathbb{C}}(0, \alpha_{\mathcal{J}} P)$ denotes the signal corresponding to the message $M_{\mathcal{J}}$ intended to the subset $\mathcal{J} \subseteq \{1, 2, 3\}$. User 3 (the weakest user) decodes $\tilde{M}_3 = \{M_3, M_{13}, M_{23}, M_{123}\}$ by treating all the other messages as noise. User 2 decodes first the messages $\tilde{M}_3$ and then jointly decodes $\tilde{M}_2 = \{M_2, M_{12}\}$. Finally, user 1 (the strongest user) decodes successively $\tilde{M}_3, \tilde{M}_2$ then finally M_1.

B. Weighted sum rate maximization

In order to characterize the boundary of the capacity region $\Gamma(\boldsymbol{h})$, we consider the weighted sum rate maximization given as

$$\max_{\boldsymbol{r} \in \Gamma(\boldsymbol{h})} \sum_{\mathcal{J}: \mathcal{J} \subseteq [K]} \theta_{\mathcal{J}} r_{\mathcal{J}}. \tag{9}$$

By exploiting a simple property of the capacity region, the problem at hand can be cast into a simpler problem as summarized below.

Theorem 2. *The weighted sum rate maximization with $2^K - 1$ variables in* (9) *reduces to a simpler problem with K variables, given by*

$$f(\boldsymbol{\alpha}) = \sum_{k=1}^{K} \phi_k \log \frac{1 + h_k \sum_{j=1}^{k} \alpha_j P}{1 + h_k \sum_{j=1}^{k-1} \alpha_j P},$$

where ϕ_k denotes the largest weight for user k

$$\phi_k \triangleq \max_{\mathcal{K}: k \in \mathcal{K} \subseteq [k]} \theta_{\mathcal{K}}.$$

Proof: The proof builds on the simple structure of the capacity region. We first remark that for a given power allocation of other users, user k sees 2^{k-1} messages $\{M_{\mathcal{J}}\}$ for $k \in \mathcal{J} \subseteq [k]$ with the equal channel gain. For a given power allocation α^k, the capacity region of these messages is a simple hyperplane characterized by 2^{k-1} vertices $C_k \boldsymbol{e}_i$ for $i = 1, \ldots, 2^{k-1}$, where C_k is the sum rate of user k in the RHS of (8) and $\boldsymbol{e}_i$ is a vector with one for the i-th entry and zero for the others. Therefore, the weighted sum rate is maximized for user k by selecting the vertex corresponding to the largest weight, denoted by ϕ. This holds for any k. ∎

We provide an efficient algorithm to solve this power allocation problem as a special case of the parallel Gaussian broadcast channel studied in [13, Theorem 3.2]. Following [13], we define the rate utility function for user k given by

$$u_k(z) = \frac{\phi_k}{1/h_k + z} - \lambda,$$

where λ is a Lagrange multiplier. The optimal solution corresponds to selecting the user with the maximum rate utility at each z and the resulting power allocation for user k is given as

$$\alpha_k^* = \left\{ z : [\max_j u_j(z)]_+ = u_k(z) \right\} / P, \tag{10}$$

with λ satisfying $P = \left[\max_k \frac{\phi_k}{\lambda} - \frac{1}{h_k} \right]_+$.

C. Application example

In this subsection, we consider the long-term average sum content delivery maximization as one of the applications of the weighted sum rate maximization solved previously. By treating a codeword intended to a subset $\mathcal{K}$ of users as a message intended to the same subset, i.e. $M_{\mathcal{K}} = V_{\mathcal{K}}$ in (2) and assuming that these codewords for different subsets are all independent, the sum content delivery rate achieved by superposition coding can be written as the weighted sum rate:

$$\sum_{\mathcal{K}: \mathcal{K} \subseteq [K]} \theta_{\mathcal{K}} R_{\mathcal{K}} \quad \text{with} \quad \theta_{\mathcal{K}} = \frac{|\mathcal{K}|}{T(m, |\mathcal{K}|)},$$

where $R_{\mathcal{K}}$ denotes the rate of message $M_{\mathcal{K}}$ satisfying the constraints in Theorem 1. By noting that the weights depend only on the cardinality of $\mathcal{K}$ and that the function $k/T(m,k)$ is increasing in k, we have the following properties i) $\theta_{\mathcal{K}} = \theta_{\mathcal{K}'}, \ \forall \mathcal{K}, \mathcal{K}'$ such that $|\mathcal{K}| = |\mathcal{K}'|$, ii) $\theta_{\mathcal{K}} < \theta_{\mathcal{J}}, \ \forall \mathcal{K} \subset \mathcal{J}$.

These properties readily imply that the effective weight of user k, denoted by ϕ_k, is given by

$$\phi_k = \max_{\mathcal{J}: k \in \mathcal{J} \subseteq [k]} \theta_{\mathcal{J}} = \frac{k}{T(m,k)}.$$

Following Theorem 2, the resulting sum delivery rate of superposition coding for a given channel state such that $h_1 \geq \cdots \geq h_K$ is given by

$$R_{\text{sum,sp}}(\boldsymbol{h}) = \sum_{k=1}^{K} \frac{k}{T(m,k)} \log \left(1 + \frac{h_k \alpha_k^* P}{1 + h_k \sum_{j=1}^{k-1} \alpha_j^* P} \right),$$

where $\{\alpha_j^*\}$ is the optimal power allocation in (10). The long-term average sum delivery rate is given by

$$R_{\text{sum,sp}} = \mathbb{E}_{\boldsymbol{h}}[R_{\text{sum,sp}}(\boldsymbol{h})].$$

IV. Performance Analysis

In this section, we analyze the long-term average sum delivery rate of the proposed scheme in two cases of interest: a) a user selection scheme that selects the best subset of users as a function of the channel state and the weights, b) naive coded caching (or baseline scheme) that applies coded caching to K users as described in Section II. By restricting ourselves to the symmetric fading case ($\gamma_k = 1, \forall k$), we consider two regimes of interest, i.e. large K and high SNR.

A. Baseline scheme: naive coded caching

In this scheme, the server serves all K users with the multicast rate limited by the worst user as in (5). We define the exponential integral function $E_1(x) = \int_1^{+\infty} \frac{e^{-xt}}{t} dt$. The performance of this scheme is summarized below.

Proposition 1. *(i)* $R_{\mathrm{sum,bl}}(K,P) = \phi_K e^{\frac{K}{P}} E_1\left(\frac{K}{P}\right)$.
(ii) For all P: $R_{\mathrm{sum,bl}}(K,P) \sim \frac{Pm}{1-m}$ *when* $K \to \infty$.
(iii) For all K: $R_{\mathrm{sum,bl}}(K,P) \sim \phi_K \log(P)$ *when* $P \to \infty$.

Proof: See Appendix VII-B. ∎

B. User selection scheme: opportunistic scheduling

Albeit suboptimal, we consider a simple time-sharing strategy, which allocates a fraction of time $\eta_{\mathcal{K}}$ to the subset of users $\mathcal{K}$, with $\sum_{\mathcal{K}\subseteq[K]} \eta_{\mathcal{K}} = 1$. The corresponding weighted sum rate maximization is given by

$$\max_{\boldsymbol{\eta}:\sum_{\mathcal{K}} \eta_{\mathcal{K}}=1} \sum_{\mathcal{K}\subseteq[K]} \theta_{\mathcal{K}} \eta_{\mathcal{K}} \log(1 + P \min_{k\in\mathcal{K}} h_k).$$

Let $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ denote the permutation such that $h_{\pi_1} \geq \cdots \geq h_{\pi_K}$. Because of the capacity region structure, the problem at hand can be simplified into:

$$\max_{\boldsymbol{\eta}} \sum_{k=1}^{K} \phi_k \eta_k \log(1 + h_{\pi_k} P).$$

The optimal solution is readily given by

$$\eta_k = \begin{cases} 1, & \text{if } k = \arg\max_j \phi_j \log(1 + h_{\pi_j} P), \\ 0, & \text{otherwise.} \end{cases}$$

This means that we transmit to only one set of users maximizing the instantaneous weighted rate with full power. By transmitting opportunistically to the group of users with the highest sum content delivery rate at each channel realization, the long-term average sum content delivery rate is given by

$$R_{\mathrm{sum,sc}}(K,P) = \mathbb{E}\left[\max_k \phi_k \log(1 + h_{\pi_k} P)\right].$$

We characterize $R_{\mathrm{sum,sc}}(K,P)$ in two regimes of interest.

Proposition 2. *(i) For all* P:
$R_{\mathrm{sum,sc}}(K,P) \sim \frac{Km}{1-m} e^{(\frac{1}{P} - \frac{1}{W(P)})} W(P)$ *when* $K \to \infty$, *where* $W(x)$ *is the Lambert function i.e.* $W(x)e^{W(x)} = x$.
(ii) For all K: $R_{\mathrm{sum,sc}}(K,P) \sim \phi_K \log(P)$ *when* $P \to \infty$.

Proof: See Appendix VII-C. ∎

C. Interpretation of the results

From propositions 1 and 2, the following remarks are in order: 1) in the large K regime, the long-term sum delivery rate of the selection scheme grows linearly for any finite SNR. This is in a sharp contrast with the baseline scheme, whose sum delivery rate converges to a constant; 2) in the high SNR regime, both schemes yield the same performance, i.e. $\frac{K}{T(m,K)} \log P$, for any finite K because the sum delivery rate is no longer sensitive to the randomness of channels and is maximized solely by exploiting the global caching gain; 3) It is worth noticing that the performance of selection scheme can be achieved without instantaneous channel knowledge. Namely, each user can measure its SNR and send a one-bit feedback indicating whether it is above or below the threshold value given by $Pz^* = \frac{P}{W(P)} - 1$.

V. Numerical Examples

In this section, we compare our proposed superposition scheme, its two special cases (baseline and selection), as well as uncoded caching. Uncoded caching refers to the case where the server sends the remaining $(1-m)F$ bits of the requested file at rate $\log(1+Ph_k)$ for each user k. Thus, the corresponding long-term average sum delivery rate is given by

$$\mathbb{E}\left[K\left(\sum_{k=1}^{K} \frac{1-m}{\log(1+Ph_k)}\right)^{-1}\right].$$

We consider a database of size $N = 10^4$, normalized memory size of $m = 10^{-1}$. In Fig. 1, we plot the long-term sum content delivery rate as a function of the number of users at $P = 10$ dB for both centralized (dashed line) and decentralized (solid line) placement strategies. We observe that both the superposition schemes and the selection scheme offer a linear increase, whereas the performance of baseline and uncoded schemes is bounded. This behavior agrees with the analysis of the previous section and implies that the performance of coded caching at low to moderate SNR is limited by the vanishing multicast rate. Furthermore, the selection scheme offers performance almost as good as the superposition scheme, despite its reduced complexity.

In Fig. 2, the long-term average sum content delivery rate is plotted as a function of SNR for different schemes. We observe that the performance of selection, baseline scheme becomes identical for large SNR, which confirms our analysis. In addition, the sum content delivery rate increases as SNR with a prelog of ϕ_K, which in turn depends on the placement strategy (3). By comparing uncoded caching and the baseline scheme, we observe that after a certain SNR threshold, the baseline scheme performs better than uncoded caching scheme.

VI. Conclusion

We have studied content delivery using coded caching over fading broadcast channels. Contrary to the baseline scheme applying coded caching to K users irrespectively of channel state information, we proposed opportunistic delivery schemes that achieve a linear increase of the sum content delivery rate by a careful selection of the user subset as a function of both channel state information and priorities. In order to reduce the amount and accuracy of feedback, we proposed a simple threshold-based feedback scheme yielding the same scalable solution while requiring only one bit per user. In future work, we plan on providing a detailed analysis of the performance of the more general superposition scheme proposed here.

References

[1] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

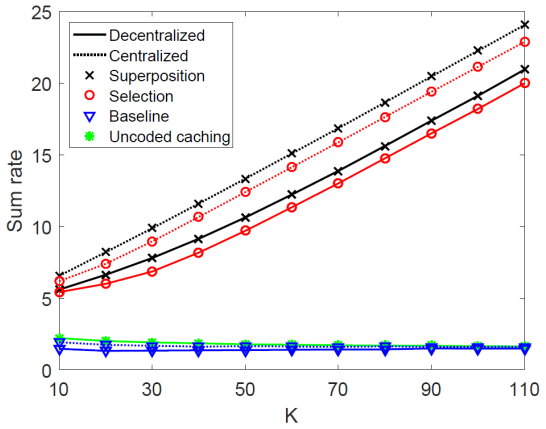

Fig. 1. Sum rate vs K for $P = 10$(dB).

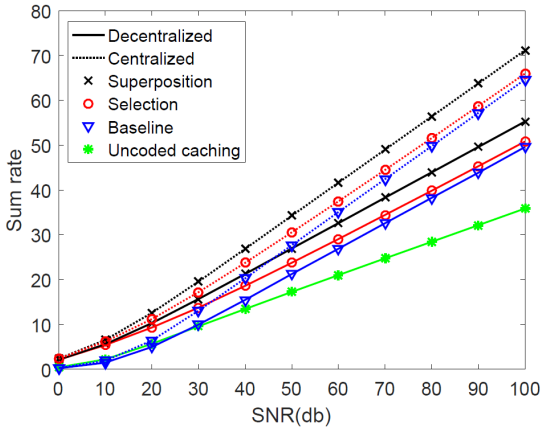

Fig. 2. Sum rate vs SNR for $K = 10$.

[2] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, G. Caire, "FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers" , *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, 2013.

[3] M. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff", *IEEE/ACM Trans. on Networking*, vol. 23, no. 4, pp. 1029–1040, 2015. .

[4] G. S. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers", *IEEE Communications Magazine*, 2016.

[5] S. S. Bidokhti, M. Wigger, and R. Timo, "Noisy Broadcast Networks with Receiver Caching", arXiv preprint arXiv:1605.02317, 2016.

[6] J. Zhang, and P. Elia, "Wireless Coded Caching: a Topological Perspective". arXiv:1606.08253, 2016.

[7] K-H. Ngo, S. Yang, and M. Kobayashi, "Cache-Aided Content Delivery in MIMO Channels", in *Proc. Allerton*, IL, USA, 2016.

[8] 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, 2016.

[9] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Multi-Antenna Coded Caching", arXiv preprints, arXiv:1701.02979, 2017.

[10] N. Jindal and Z. Q. Luo, "Capacity limits of multiple antenna multicast", *2006 IEEE International Symposium on Information Theory*, 2006, pp. 1841–1845

[11] A. Destounis, M. Kobayashi, G. Paschos, and A. Ghorbel, "Alpha Fair Coded Caching", arXiv preprint, arXiv:1701.07730.

[12] A. El Gamal and Y-H. Kim, "Network information theory", *Cambridge university press*, 2011.

[13] D. N. Tse, "Optimal power allocation over parallel Gaussian broadcast channels", *Electronics Research Laboratory, College of Engineering, University of California*, 1999.

VII. Appendix

A. Proof of Theorem 1

We provide the proof for $K = 3$ and the general case $K > 3$ follows readily. Here log denotes the binary logarithm, $h(.)$ denotes the differential entropy, and $H(.)$ denotes Shannon entropy.

Converse Notice that the channel output of user k in (4) for n channel use can be equivalently written as

$$y_{k,i} = x_i + \tilde{w}_{k,i}, \qquad i = 1, \ldots, n \tag{11}$$

where $\tilde{w}_{k,i} = \frac{w_k[i]}{\sqrt{h_k}} \sim \mathcal{N}_{\mathbb{C}}(0, N_k)$ for $N_k = \frac{1}{h_k}$. Since $N_1 \leq N_2 \leq N_3$, we set $\tilde{M}_k = \cup_{k \in \mathcal{J} \subseteq [k]} M_{\mathcal{J}}$ the message set that must be decoded by user k at sum rate $\tilde{R}_k = \cup_{k \in \mathcal{J} \subseteq [k]} R_{\mathcal{J}}$. More explicitly, we have $\tilde{M}_1 = \{M_1\}$, $\tilde{M}_2 = \{M_2, M_{12}\}$, $\tilde{M}_3 = \{M_3, M_{13}, M_{23}, M_{123}\}$. By Fano's inequality, we have

$$\begin{cases} nH(\tilde{M}_1) & \leq I(\tilde{M}_1; Y_1 \,|\, \tilde{M}_2, \tilde{M}_3) \\ nH(\tilde{M}_2) & \leq I(\tilde{M}_2; Y_2 \,|\, \tilde{M}_3) \\ nH(\tilde{M}_3) & \leq I(\tilde{M}_3; Y_3). \end{cases} \tag{12}$$

Consider first user 3.

$$I(\tilde{M}_3; Y_3) = h(Y_3) - h(Y_3 \,|\, \tilde{M}_3). \tag{13}$$

Since we have $n \log(\pi e N_3) = h(Y_3 \,|\, \tilde{M}_3, X) \leq h(Y_3 \,|\, \tilde{M}_3) \leq h(Y_3) \leq n \log(\pi e(P + N_3))$, there exists $0 \leq \alpha_3 \leq 1$ such that

$$h(Y_3 \,|\, \tilde{M}_3) = n \log(\pi e((1 - \alpha_3)P + N_3)). \tag{14}$$

Using (13) and (14) we obtain

$$\begin{aligned} &I(\tilde{M}_3; Y_3) \\ &\quad = h(Y_3) - h(Y_3 \,|\, \tilde{M}_3) \\ &\quad \leq n \log(\pi e(P + N_3)) - n \log(\pi e((1 - \alpha_3)P + N_3)) \\ &\quad = n \log\left(\frac{N_3 + P}{N_3 + (1 - \alpha_3)P}\right). \end{aligned} \tag{15}$$

Next consider user 2.

$$I(\tilde{M}_2; Y_2 \,|\, \tilde{M}_3) = h(Y_2 \,|\, \tilde{M}_3) - h(Y_2 \,|\, \tilde{M}_2, \tilde{M}_3). \tag{16}$$

Using the conditional entropy power inequality in [12] , we have

$$\begin{aligned} h(Y_3 \,|\, \tilde{M}_3) &= h(Y_2 + \tilde{W}_3 - \tilde{W}_2 \,|\, \tilde{M}_3) \\ &\geq n \log(2^{h(Y_2 \,|\, \tilde{M}_3)/n} + 2^{h(\tilde{W}_3 - \tilde{W}_2 \,|\, \tilde{M}_3)/n}) \\ &= n \log(2^{h(Y_2 \,|\, \tilde{M}_3)/n} + \pi e(N_3 - N_2)). \end{aligned} \tag{17}$$

(14) and (17) imply

$$\begin{aligned} &n \log(\pi e((1 - \alpha_3)P + N_3)) \\ &\qquad \geq n \log(2^{h(Y_2 \,|\, \tilde{M}_3)/n} + \pi e(N_3 - N_2)) \end{aligned}$$

equivalent to

$$h(Y_2 \,|\, \tilde{M}_3) \leq n \log(\pi e((1 - \alpha_3)P + N_2)). \tag{18}$$

Since $n \log(\pi e N_2) = h(Y_2 \,|\, \tilde{M}_2, \tilde{M}_3, X) \leq h(Y_2 \,|\, \tilde{M}_2, \tilde{M}_3) \leq h(Y_2 \,|\, \tilde{M}_3) \leq n \log(\pi e((1 - \alpha_3)P + N_2))$, there exists α_2 such that $0 \leq 1 - \alpha_2 - \alpha_3 \leq 1 - \alpha_3$ and

$$h(Y_2 \,|\, \tilde{M}_2, \tilde{M}_3) = n \log(\pi e((1 - \alpha_2 - \alpha_3)P + N_2)). \tag{19}$$

Using (16), (18) and (19) it follows

$$\begin{aligned} I(\tilde{M}_2; Y_2 \,|\, \tilde{M}_3) &= h(Y_2 \,|\, \tilde{M}_3) - h(Y_2 \,|\, \tilde{M}_2, \tilde{M}_3) \\ &\le n \log(\pi e((1-\alpha_3)P + N_2)) \\ &\quad - n \log(\pi e((1-\alpha_2-\alpha_3)P + N_2)) \\ &= n \log\left(\frac{N_2 + (1-\alpha_3)P}{N_2 + (1-\alpha_2-\alpha_3)P}\right). \end{aligned} \tag{20}$$

Finally we consider user 1.

$$\begin{aligned} &I(\tilde{M}_1; Y_1 \,|\, \tilde{M}_2, \tilde{M}_3) \\ &\quad = h(Y_1 \,|\, \tilde{M}_2, \tilde{M}_3) - h(Y_1 \,|\, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3) \\ &\quad \le h(Y_1 \,|\, \tilde{M}_2, \tilde{M}_3) - h(Y_1 \,|\, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3, X) \\ &\quad = h(Y_1 \,|\, \tilde{M}_2, \tilde{M}_3) - h(Y_1 \,|\, X) \qquad (21) \\ &\quad = h(Y_1 \,|\, \tilde{M}_2, \tilde{M}_3) - n \log(\pi e N_1) \qquad (22) \end{aligned}$$

where (21) holds because $\left(\tilde{M}_1, \tilde{M}_2, \tilde{M}_3\right) \to X \to Y_1$ is a Markov chain. Using the conditional entropy power inequality in [12] , we have

$$\begin{aligned} &h(Y_2 \,|\, \tilde{M}_2, \tilde{M}_3) \\ &\quad = h(Y_1 + \tilde{W}_2 - \tilde{W}_1 \,|\, \tilde{M}_2, \tilde{M}_3) \\ &\quad \ge n \log(2^{h(Y_1 \,|\, \tilde{M}_2, \tilde{M}_3)/n} + 2^{h(\tilde{W}_2 - \tilde{W}_1 \,|\, \tilde{M}_2, \tilde{M}_3)/n}) \\ &\quad = n \log(2^{h(Y_1 \,|\, \tilde{M}_2, \tilde{M}_3)/n} + \pi e(N_2 - N_1)) \qquad (23) \end{aligned}$$

(19) and (23) imply

$$\begin{aligned} &n \log(\pi e((1-\alpha_2-\alpha_3)P + N_2)) \\ &\quad \ge n \log(2^{h(Y_1 \,|\, \tilde{M}_2, \tilde{M}_3)/n} + \pi e(N_2 - N_1)) \end{aligned}$$

equivalent to

$$h(Y_1 \,|\, \tilde{M}_2, \tilde{M}_3) \le n \log(\pi e((1-\alpha_2-\alpha_3)P + N_1)). \tag{24}$$

Let $\alpha_1 = 1 - \alpha_2 - \alpha_3$. Combining the last inequality with (22) we obtain

$$I(\tilde{M}_1; Y_1 \,|\, \tilde{M}_2, \tilde{M}_3) \le n \log\left(\frac{N_1 + \alpha_1 P}{N_1}\right). \tag{25}$$

From (12), (15), (20) and (25), it readily follows that $\exists\ 0 \le \alpha_1, \alpha_2, \alpha_3 \le 1$ such that $\alpha_1 + \alpha_2 + \alpha_3 = 1$ and

$$\begin{cases} H(\tilde{M}_1) & \le \log\left(1 + \frac{\alpha_1 P}{N_1}\right), \\ H(\tilde{M}_2) & \le \log\left(1 + \frac{\alpha_2 P}{N_2 + \alpha_1 P}\right), \\ H(\tilde{M}_3) & \le \log\left(1 + \frac{\alpha_3 P}{N_3 + (\alpha_1+\alpha_2)P}\right). \end{cases}$$

By replacing $H(\tilde{M}_k)$ with $\sum_{k \in \mathcal{K} \subseteq [k]} R_{\mathcal{K}}$ and N_k with $\frac{1}{h_k}$ we obtain the result

$$\begin{cases} R_1 & \le \log(1 + h_1 \alpha_1 P) \\ R_2 + R_{12} & \le \log\left(\frac{1 + h_2(\alpha_1+\alpha_2)P}{1 + h_2 \alpha_1 P}\right) \\ R_3 + R_{13} + R_{23} + R_{123} & \le \log\left(\frac{1 + h_3 P}{1 + h_3(\alpha_1+\alpha_2)P}\right), \end{cases}$$

Achievability We prove that superposition coding achieves the upper bound. For $\mathcal{J} \subseteq \{1,2,3\}$, generate random sequences $x_{\mathcal{J}}(m_{\mathcal{J}})$, $m_{\mathcal{J}} \in [1 : 2^{nR_{\mathcal{J}}}]$ each i.i.d. $\mathcal{N}_{\mathbb{C}}(0, \alpha_{\mathcal{J}} P)$, where $\sum_{\mathcal{J} \subseteq \{1,2,3\}} \alpha_{\mathcal{J}} = 1$. We define $\tilde{x}_k(\tilde{m}_k) = \sum_{k \in \mathcal{J} \subseteq [k]} x_{\mathcal{J}}(m_{\mathcal{J}})$, where $\tilde{m}_k \in [1 : 2^{n\tilde{R}_k}]$. To transmit $\{m_{\mathcal{J}}\}_{\mathcal{J} \subseteq \{1,2,3\}}$, the encoder set $X = \sum_{\mathcal{J} \subseteq \{1,2,3\}} x_{\mathcal{J}}(m_{\mathcal{J}}) = \tilde{x}_1(\tilde{m}_1) + \tilde{x}_2(\tilde{m}_2) + \tilde{x}_3(\tilde{m}_3)$. For decoding:

- Receiver 3 jointly decodes $\{m_3, m_{13}, m_{23}, m_{123}\}$ by treating $\tilde{x}_1(\tilde{m}_1)$ and $\tilde{x}_2(\tilde{m}_2)$ as noise.
- Receiver 2 uses successive cancellation by first decoding $\tilde{x}_3(\tilde{m}_3)$ and treating $\tilde{x}_1(\tilde{m}_1)$ and $\tilde{x}_2(\tilde{m}_2)$ as noise. It recovers $\{m_{23}, m_{123}\}$. By subtracting off $\tilde{x}_3(\tilde{m}_3)$ and treating $\tilde{x}_1(\tilde{m}_1)$ as noise, user 2 decodes $\tilde{x}_2(\tilde{m}_2)$ from which it recovers $\{m_2, m_{12}\}$.
- Receiver 1 decodes $\tilde{x}_3(\tilde{m}_3)$ and recovers $\{m_{13}, m_{123}\}$. Then, by successive cancellation it decodes $\tilde{x}_2(\tilde{m}_2)$ and recovers $\{m_{12}\}$. Finally it decodes $\tilde{x}_1(\tilde{m}_1)$ to recover $\{m_1\}$.

B. Proof of Proposition 1

The content delivery rate is:

$$R_{\mathrm{sum,bl}}(K, P) = \phi_K \mathbb{E}\left[\log\left(1 + P h_{\min}\right)\right],$$

where $h_{\min} \triangleq \min_{k=1,\dots,K} h_k$. Since $(h_k)_{k=1,\dots,K}$ are i.i.d. with distribution Exp(1), $h_{\min}$ has distribution Exp(K). Hence:

$$\begin{aligned} \mathbb{E}\left[\log\left(1 + P h_{\min}\right)\right] &= \int_0^{+\infty} e^{-x} \log\left(1 + \frac{P}{K} x\right) dx \\ &= e^{\frac{K}{P}} E_1\left(\frac{K}{P}\right), \end{aligned}$$

which yields statement (i).

When $K \to \infty$ we have $\phi_K \sim \frac{Km}{1-m}$ and

$$\int_0^{+\infty} e^{-x} \log\left(1 + \frac{P}{K} x\right) dx \sim \frac{P}{K} \int_0^{+\infty} x e^{-x} dx = \frac{P}{K},$$

Replacing yields statement (ii):

$$R_{\mathrm{sum,bl}}(K, P) \sim \frac{Pm}{1-m}.$$

When $P \to \infty$, $\frac{K}{P} \to 0$. Since $E_1(x) \sim \log(1/x)$ for $x \to 0$ we obtain statement (iii):

$$R_{\mathrm{sum,bl}}(K, P) \sim \phi_K \log(P/K) \sim \phi_K \log(P).$$

C. Proof of proposition 2

We start by statement (i). The proof involves upper and lower bounding $R_{\mathrm{sum,sc}}(K, P)$ by two expressions which are equivalent in the large K regime. We define the complementary c.d.f. of $(h_k)_{k=1,\dots,K}$:

$$U(z) \triangleq \sum_{k=1}^{K} \mathbf{1}\{h_k \ge z\},$$

with $z \ge 0$. We further define the function:

$$g(z) \triangleq \frac{m}{1-m} e^{-z} \log(1 + Pz), z \ge 0.$$

It is noted that $g(0) = g(\infty) = 0$, and that g is smooth. Differentiating, we have that g is maximized at:

$$z^*(P) \triangleq \arg\max_{z \ge 0} g(z) = \frac{1}{W(P)} - \frac{1}{P}$$

so that:

$$\max_{z\geq 0} g(z) = g(z^*) = \frac{m}{1-m} e^{(\frac{1}{P} - \frac{1}{W(P)})} W(P).$$

The proof relies on the following equality:

$$\begin{aligned}\max_{k=1,\dots,K} \phi_k \log(1+h_{\pi_k}P) &= \max_{z\in\{h_1,\dots,h_K\}} \phi_{U(z)} \log(1+Pz) \\ &= \max_{z\geq 0} \phi_{U(z)} \log(1+Pz).\end{aligned}$$

Indeed, function $z \mapsto \phi_{U(z)} \log(1+Pz)$ is left-continuous and is both continuous and increasing for all $z \notin \{h_1, ..., h_K\}$ so that it must attain its maximum in the set $\{h_1, ..., h_K\}$.

Lower bound Using the previous equality we obtain:

$$\begin{aligned}R_{\mathrm{sum,sc}}(K,P) &= \mathbb{E}\left[\max_{z\geq 0} \phi_{U(z)} \log(1+Pz)\right] \\ &\geq \max_{z\geq 0} \mathbb{E}\left[\phi_{U(z)}\right] \log(1+Pz) \quad (26) \\ &\geq \max_{z\geq 0} \frac{m}{1-m} \mathbb{E}\left[U(z)\right] \log(1+Pz) \quad (27) \\ &= \max_{z\geq 0} \frac{m}{1-m} K e^{-z} \log(1+Pz) \quad (28) \\ &= K \max_{z\geq 0} g(z) \\ &= \frac{Km}{1-m} e^{(\frac{1}{P} - \frac{1}{W(P)})} W(P), \quad (29)\end{aligned}$$

where (26) follows from Jensen's inequality; (27) from the fact that $\phi_k \geq \frac{mk}{1-m}$ and (28) from $\mathbb{E}(U(z)) = Ke^{-z}$.

Upper bound The upper bound is slightly more involved and involves a dominated convergence argument. Let us define:

$$G(K) = \frac{1}{K} \max_{z\geq 0} \phi_{U(z)} \log(1+Pz),$$

so that $R_{\mathrm{sum,sc}}(K,P) = K\mathbb{E}(G(K))$. We prove that:

(a) $\sup_K \mathbb{E}(G(K)) < \infty$ and

(b) $\limsup_{K\to\infty} G(K) \overset{a.s.}{\leq} g(z^*)$

If both (a) and (b) holds, applying the reverse Fatou lemma proves the announced result:

$$\limsup_{K\to\infty} \frac{R_{\mathrm{sum,sc}}(K,P)}{K} = \limsup_{K\to\infty} \mathbb{E}(G(K)) \leq g(z^*).$$

Consider claim (a). Since $\phi_k \leq k \ \forall k$:

$$\begin{aligned}\phi_{U(z)} \log(1+Pz) &\leq U(z) \log(1+Pz) \\ &= \sum_{k=1}^{K} \mathbf{1}\{h_k \geq z\} \log(1+Pz) \\ &\leq \sum_{k=1}^{K} \log(1+Ph_k).\end{aligned}$$

The above holds for all z, and taking expectations:

$$\begin{aligned}\mathbb{E}(G(K)) &= \frac{1}{K} \mathbb{E}(\sup_{z\geq 0} \phi_{U(z)} \log(1+Pz)) \\ &\leq \mathbb{E}(\log(1+Ph_k) < \infty.\end{aligned}$$

The above holds for all K, so that $\sup_K \mathbb{E}(G(K)) < \infty$.

We turn to claim (b). Consider $y > z^*(P)$ fixed, whose value will be made precise afterwards. Define intervals $I_0 \triangleq [0,y]$, $I_1 \triangleq [y,\infty)$ and for $i \in \{0,1\}$, define:

$$G_i(K) = \frac{1}{K} \max_{z\in I_i} \{\phi_{U(z)} \log(1+Pz)\},$$

so that $G(K) = \max\{G_0(K), G_1(K)\}$. To prove that $\limsup_{K\to\infty} G(K) \leq g(z^*)$ it is sufficient to prove that $\limsup_{K\to\infty} G_i(K) \leq g(z^*)$ for $i \in \{0,1\}$.

Consider $G_0(K)$. For $z \in I_0$, we have $U(z) \geq U(y)$, so that:

$$\phi_{U(z)} = \frac{U(z)}{T(m,U(z))} \leq \frac{U(z)}{T(m,U(y))}.$$

Therefore:

$$G_0(K) \leq \frac{1}{T(m,U(y))} \max_{z\in I_0} \left\{\frac{U(z)}{K} \log(1+Pz)\right\}. \quad (30)$$

The Glivenko-Cantelli theorem states that:

$$\sup_{z\geq 0} \left|\frac{U(z)}{K} - e^{-z}\right| \xrightarrow{a.s.} 0$$

so that:

$$\begin{aligned}&\max_{z\in I_0} \left|\frac{U(z)}{K} \log(1+Pz) - e^{-z}\log(1+Pz)\right| \\ &\leq \max_{z\geq 0} \left|\frac{U(z)}{K} - e^{-z}\right| \log(1+Py) \xrightarrow{a.s.} 0 \quad (31)\end{aligned}$$

From the law of large numbers $U(y) \xrightarrow{a.s.} \infty$, so $T(m,U(y)) \xrightarrow{a.s.} \frac{1-m}{m}$, together with (30) and (31) it implies

$$\lim_{K\to\infty} \sup G_0(K) \overset{a.s.}{\leq} \max_{0\leq z\leq y} g(z). \quad (32)$$

Now, consider $G_1(K)$. For $z \in I_1$, by the same argument as previously:

$$\begin{aligned}\frac{1}{K}\phi_{U(z)} \log(1+Pz) &\leq \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\{h_k \geq z\} \log(1+Ph_k) \\ &\leq \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\{h_k \geq y\} \log(1+Ph_k) \\ &\overset{a.s.}{\to} \mathbb{E}(\mathbf{1}\{h_k \geq y\} \log(1+Ph_k)),\end{aligned}$$

using the law of large numbers.

Since $y \to \mathbb{E}(\mathbf{1}\{h_k \geq y\} \log(1+Ph_k))$ is decreasing and vanishes when $y \to \infty$, we may select y large enough so that:

$$\mathbb{E}(\mathbf{1}\{h_k \geq y\} \log(1+Ph_k)) \leq g(z^*).$$

Putting it together $\limsup G_1(K) \overset{a.s.}{\leq} g(z^*)$ which is claim (b). This concludes the proof of statement (i).

Consider statment (ii), we have for $P \to \infty$:

$$\frac{\max_k \phi_k \log(1+h_{\pi_k}P)}{\log(P)} \overset{a.s.}{\to} \phi_K.$$

Furthermore,

$$\sup_{P\geq 0} \mathbb{E}\left(\frac{\max_k \phi_k \log(1+h_{\pi_k}P)}{\log(P)}\right) = \sup_{P\geq 0} \frac{R_{\mathrm{sum,sc}}(K,P)}{\log(P)} < \infty$$

so by Lebesgue's theorem $\frac{R_{\mathrm{sum,sc}}(K,P)}{\log(P)} \to \phi_K$.