Optimal Power Allocation in MISO Cache-Aided Communication

Soheil Mohajer
Department of ECE
University of Minnesota
Email: soheil@umn.edu

Itsik Bergel
Faculty of Engineering
Bar-Ilan University
Email: itsik.bergel@biu.ac.il

***Abstract*—The problem of cache-aided communication is studied for cellular networks with multi-user, with multiple-input and multiple output antennas, and finite range of signal-to-noise ratio. Users are assumed to have non-symmetric links, modeled by wideband fading channels. The ultimate goal is to design a joint power, rate, and cache allocation along with pre-fetching and fetching schemes to minimize the duration of communication in the delivery phase. This is formulated as a linear program, and a closed form solution is presented for some special range of parameters. In particular, the gain of MIMO coded caching is shown to be beyond trading between spatial diversity and accumulative cache size, specially in low SNR regime, where each user can be served at its own channel rate, rather than being compromised by the poorest user in the communication group.**

***Index Terms*—Cache-aided communication, MIMO, Finite SNR regime, Cache and Power allocation, Zero-forcing.**

I. Introduction

Network traffic has experienced dramatic growth in recent years. Video delivery accounts for a major growth of traffic on mobile networks. In video content delivery popular files are repeatedly requested by multiple users. Furthermore, video request has a prime time. These properties provide an opportunity for storing the data at local caches during the off-peak hours to serve the requests at the peak hours.

Data caching refers to bringing the data closer to where it will be used. In cellular networks, however, the available cache at users is limited and typically negligible compared to the size of popular data that might be requested by the user. In a pioneering work [1], it was shown that caching gain is not limited to the local cache size at individual users. The proposed coded caching scheme offers a throughput gain which scales with the aggregate size of the caches distributed across all users in the network. This scheme is based on the fact that caching a packet at one user provides an opportunity for *multicasting combined packets*, even if it is only requested by another user. This scheme was shown to be information-theoretically optimal [2], under ideal assumptions.

This scheme was further generalized to multiple transmit antennas [3], where it was shown that a network can achieve $N+M$ degrees of freedom (DoF), where N is the number of antennas and M is the number of copies of the complete dataset stored across all users. The potential gain is substantial even in spite of the current trend of massive MIMO, especially since a small cache at each mobile comes with very low cost, and these memories easily scale up with the number of users, that can offer significant gains.

The cache aided communication problem has been studied further towards their adoption in wireless networks. In particular, [4]–[6] studied the effect of fading and/or erasure channels, [6], [7] extended the approach to multiple antennas at transmitters and/or receivers, and [8] considered single antenna network with heterogeneous rates. Nevertheless, existing works in this area only focus on isolated scenarios, or are limited to DoF characterization (i.e., asymptotically high SNR). Thus, there is a big gap to be covered before we can fully understand and exploit the role of cache-enabled communication in cellular networks.

In practice, users are located at different distances and are subject to power attenuation, and fading. The gain of caching and optimum caching strategies in such a multi-antenna scenario are still unknown. Employment of coded caching in wireless networks also requires addressing the resource allocation issues. Several recent works considered the cache allocation problem [9]–[11]. The focus of this work is on the power allocation, which is typically treated as a separate optimization problem, isolated from the cache aided communication. In particular, the power allocation for superposition coding in a cache-aided broadcast channel is studied in [12]–[14]. Nevertheless, the power allocation has a crucial role in determining the user rates, and hence the performance of cache-aided communication.

We focus on the scenario where the number of users exactly equals $N+M$. In spite of its limitation, to the best of our knowledge this is currently the only scenario for which a closed-form solution for optimum cache allocation and placement for non-asymptotic SNR is known [15]. We formulate the power allocation as an optimization problem, and give a simple algorithm to solve it. We show that this algorithm degenerates to the well known water-filling algorithm. As a result, the strict fairness is achieved through maximizing the total network throughput. This reveals a new phenomenon in cache aided communication: caching allows a natural balancing of the load between the users and thus enables operation with both maximal throughput and strict fairness.

The work of S. Mohajer was supported in part by the National Science Foundation under Grant CCF-1749981.

978-1-5386-3512-4/18/$31.00 ©2018 IEEE

II. System model

We consider a single cell network with one base station (BS), which has access to a dictionary of D file and serves U users, where $D \geq U$. The BS has N antennas and each user is equipped with a single antenna. We assume a wideband communication scheme, in which the bandwidth, B, is divided into F frequency bins (e.g., OFDM), and each bin $m \in \{1, \ldots, F\}$ carries one modulated symbol at a time without inter symbol interference.

The received sample after matched filtering for the m-th frequency bin at the i-th user is given by

$$y_{i,m} = \mathbf{h}_{i,m}\mathbf{x}_m + w_{i,m}, \tag{1}$$

where $\mathbf{x}_m$ is the transmit vector, $w_{i,m} \sim \mathbb{C}\mathcal{N}(0, \sigma^2)$ is the additive complex white Gaussian noise, and $\mathbf{h}_{i,m} \in \mathbb{C}^{1\times N}$ is the channel vector from the BS to User i. We assume the BS has perfect channel state information. Each link between two antennas experiences fading. Thus, the channel vector for the m-th frequency bin can be written as

$$\mathbf{h}_{i,m} = \sqrt{r_i^{-\alpha}} \cdot \mathbf{g}_{i,m} \tag{2}$$

where α is the path-loss exponent and $\mathbf{g}_{i,m} \sim \mathbb{C}\mathcal{N}(0, \mathbf{I})$ represents the random fading.

The BS simultaneously transmits different messages to each user. Transmission of some of the requested messages can be ignored if the messages are already stored in the users' cache. For other messages, the BS uses zero forcing (ZF) precoding, so that the transmission of an undesired message will not interfere with the reception of the desired user. In this work we consider a sub-optimal transmission scheme where the BS transmits each message in a way that causes no interference at all to a group of $N-1$ users. More precisely, we send

$$\mathbf{x}_m = \sum\nolimits_{i=1}^{U} \sqrt{P_i p_{i,m}}\mathbf{f}_{i,m} d_{i,m} \tag{3}$$

where $\mathbf{f}_{i,m}$ is the precoding vector (with $\|\mathbf{f}_{i,m}\|^2 = 1$), $d_{i,m}$ is the data symbol, P_i is the inter-user power allocation for user i, and $p_{i,m}$ is the intra-user power allocation for the m-th frequency bin. The intra-user power allocation is normalized to $\mathbb{E}[p_{i,m}] = 1$ for all i, and the inter-user power allocation is subject to a sum-power constraint, $\sum_i P_i \leq P$.

We focus on the case of $U = N + M$, where all users can be simultaneously served. To avoid interference, the precoding vector for each message is chosen such that it is orthogonal to the channel vectors of a selected set of $N-1$ users that should not be disturbed by this message. All other users have this undesired message in their cache, and can subtract it from their received signal. The achievable rate of user i is given by

$$\tilde{R}_i = \frac{B}{F}\sum_{m=1}^{F} \log_2\left(1 + \frac{P_i}{\sigma^2}p_{i,m}\eta_{i,m}\right), \tag{4}$$

where $\eta_{i,m} = |\mathbf{h}_{i,m}\mathbf{f}_{i,m}|^2$, and B/F is the symbol rate.

Assuming a sufficiently large bandwidth that is divided to sufficiently large number of frequency bins, F, and using the law of large numbers, the average rate per bin converges to its expectation:

$$\frac{\tilde{R}_i}{F} \to \frac{R_i}{F} = \mathbb{E}\left[\log_2\left(1 + \frac{P_i}{\sigma^2}p_{i,m}\eta_{i,m}\right)\right], \tag{5}$$

where P_i is considered deterministic, while $p_{i,m} = q(\eta_{i,m})$ is allowed to depend on the effective channel gain and hence is random. Note that from (2), $\eta_{i,m}$ admits an exponential distribution with a mean of 1.

The cache-aided communication scheme consists of two phases: placement phase and delivery phase. In the placement phase, we determine the segments of the files to be stored in the cache of user i, without knowing the users' requests. Once placement is completed, the users reveal their requests. During the delivery phase, the server broadcasts a message X such that each user can retrieve its desired file using the received message and its cache. We consider the worst case scenario, in which users request for distinct files. The goal is to minimize the duration of the delivery phase. To this end, we can optimize over power and cache allocation. Without caching, the main bottleneck is the weakest user in the network which requires maximum communication time. The general idea here is to allocate larger cache size to weaker users to compensate for their weak channel, to reduce the maximum required delay.

III. Coded Caching and Cache optimization

The placement phase of caching includes cache allocation, referring to determining the size of storage associated to each user, that is, to specify which part of each file to be pre-fetched at each user. The goal of the placement phase is to facilitate the communication in the delivery phase.

In the following, we first present the placement and delivery phase of a cache-enabled network using a simple example.

A. An Illustrative Example

Consider a MISO broadcast channel with $U = 4$ single-antenna users, and a BS with $N = 2$ antennas, as shown in Fig. 1. Each user requests one of the unit-size files from the dictionary. For the sake of illustration, we assume single band transmission, where users' links can support rate $R_1 = 5$, $R_2 = R_3 = 2$, and $R_4 = 1$ files per unit of time. We have a total cache to store $M = 2$ copies of the entire dictionary of files across the users. We denote the size of cache allocated to user u by q_u. It turns out (see Section III-B) that the optimum cache allocation is

$$q_1 = 0, \qquad q_2 = q_3 = 3/5, \qquad q_4 = 4/5. \tag{6}$$

This means, User 1 does not cache anything, while User 2 caches $3/5 = \%60$ of each file. Note that $\sum q_u = 2 = M$. After cache allocation, we have to determine the cache placement, i.e., which parts of the files are stored in the cache of each user. The optimum cache allocation and placement are demonstrated in Fig. 1. Dividing each file into 5 segments and denoting them by subscript i, for $i \in \{1, \ldots, 5\}$, User 2 caches segments 1, 2, and 3 of each file.

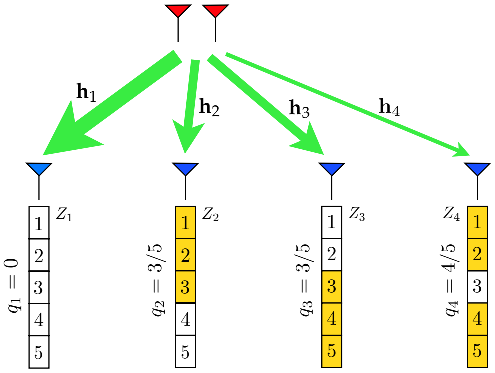

Fig. 1: A network with $N = 2$ transmit antennas, a total cache to store $M = 2$ copies of the dictionary at users, and $U = N + M = 4$ users. Cache allocation and placement are illustrated.

W.o.l.g. let users request for files A, B, C, and D, respectively. All users will be simultaneously served by broadcasting

$$X = \mathbf{h}_3^\perp A_1 + \mathbf{h}_3^\perp A_2 + \mathbf{h}_4^\perp A_3 + \mathbf{h}_2^\perp A_4 + \mathbf{h}_2^\perp A_5 + \mathbf{h}_1^\perp B_4 + \mathbf{h}_1^\perp B_5 + \mathbf{h}_1^\perp C_1 + \mathbf{h}_1^\perp C_2 + \mathbf{h}_1^\perp D_3.$$

The interference will be removed by either zero-forcing or interference suppression using the cache content of the users. Due to zero-forcing, the received signal at User 1 will be

$$Y_1 = \mathbf{h}_1 X = \mathbf{h}_1(\mathbf{h}_3^\perp A_1 + \mathbf{h}_3^\perp A_2 + \mathbf{h}_4^\perp A_3 + \mathbf{h}_2^\perp A_4 + \mathbf{h}_2^\perp A_5),$$

from which it can decode all the segments of file A. Note that this user has rate 5, and communication over $T = 1/5$ suffices to deliver 1 file. User 2, however, receives

$$Y_2 = \mathbf{h}_2 X = \mathbf{h}_2\mathbf{h}_3^\perp A_1 + \mathbf{h}_2\mathbf{h}_3^\perp A_2 + \mathbf{h}_2\mathbf{h}_4^\perp A_3 + \mathbf{h}_2\mathbf{h}_1^\perp B_4 + \mathbf{h}_2\mathbf{h}_1^\perp B_5 + \mathbf{h}_2\mathbf{h}_1^\perp C_1 + \mathbf{h}_2\mathbf{h}_1^\perp C_2 + \mathbf{h}_2\mathbf{h}_1^\perp D_3.$$

Recall that User 2 has segments 1, 2, and 3 of each file in its cache, and hence, by subtracting those, we get

$$\tilde{Y}_2 = \mathbf{h}_2\mathbf{h}_1^\perp B_4 + \mathbf{h}_2\mathbf{h}_1^\perp B_5,$$

which together with Z_2 (cache of User 2) can fulfill the request of User 2. Recall that $R_2 = 2$, and hence the duration of time for it to decode $2/5$ of a file is only $T = 1/5$. The other users can also decode their desired files in a similar manner. □

B. Cache Allocation

Our focus in this work is on the special case of $U = M+N$. Even though this is a limited range of parameters, it is an insightful analysis towards the optimal solution. In this special case, all users are active throughout all the transmissions. This makes the analysis much simpler, and allows for an analytical performance evaluation, as stated in the theorem below.

Theorem 1: For the cache-aided communication problem with $U = M + N$, if the rate of each user satisfies

$$R_u \le \frac{1}{N}\sum\nolimits_{k=1}^{U} R_k, \tag{7}$$

then the minimal time to serve all users is

$$T = \frac{N}{\sum_u R_u}. \tag{8}$$

Proof 1: For $U = M + N$ an 'efficient' transmission must serve all users. Thus the total transmission time for each user equals T. Recall that q_u fraction of the file requested by User u is pre-stored in its cache. Hence, the time required to deliver the remaining $1 - q_u$ fraction satisfies

$$1 - q_u = TR_u. \tag{9}$$

In addition, the sum of the allocations for all users satisfies

$$\sum\nolimits_{u=1}^{U} q_u = M. \tag{10}$$

Substituting (9) in (10) gives:

$$\sum\nolimits_{u=1}^{U}(1 - q_u) = U - \sum\nolimits_{u=1}^{U} q_u = T\sum\nolimits_{u=1}^{U} R_u,$$

which together with $U = M + N$ imply

$$T \ge \frac{N + M - \sum_{u=1}^{U} q_u}{\sum_{u=1}^{U} R_u} = \frac{N}{\sum_u R_u}. \tag{11}$$

However, this solution is feasible only if the resulting q_u's are positive, that is, $q_u = 1 - TR_u \ge 0$, which is equivalent to

$$R_u \le \frac{1}{T} = \frac{1}{N}\sum_{k=1}^{U} R_k. \tag{12}$$

The achievability of this result stems from the observation that this scheme is significantly simpler than other caching schemes in the sense that the transmission to each user can be optimized separately. Thus, the optimum transmission scheme always sends to *all* the users simultaneously, and interference management is performed over each individual stream: since each requested packet exists at exactly M users' cache, this users can suppress the interference using their cache content. Thus, each packet just need to be zero-forced at the $N - 1$ users that do not store this packet in their cache.

It is worth noting that in the above-mentioned example, $(R_1, R_2, R_3, R_4) = (5, 2, 2, 1)$ satisfies the constraint in (12), and hence the cache allocation obtained from (9) and given by (6) is optimal.

IV. Power Allocation

Recall from (11) that in order to minimize the transmission delay, T, it suffices to maximize the sum-rate. This could be done using the well-known water-filling algorithm. However, (11) is only valid if (12) is satisfied. That is, each individual rate should not exceed $\frac{1}{N}\sum_{k=1}^{M+N} R_k$. Intuitively, a user with $R_u > \frac{1}{N}\sum_{k=1}^{M+N} R_k$ will be fully served before the completion of serving other users. Thus, the network is left with less than $N + M$ active users for the rest of the transmission time, which cannot fully exploit the available DoF.

The control variables to maximize the sum-rate are the power shares allocated to each user u and each frequency band m, which are denoted by P_u and $p_{u,m}$, respectively. We define $Q_{u,m} = P_u \cdot p_{u,m}$ as the power allocated to user u to transmit over the m-th frequency band. Hence, the input power constraint will be $\frac{1}{B}\sum_{u,m} Q_{u,m} \le P$. The general solution for an optimum power allocation to maximize the sum-rate can be obtained by the standard water-filling algorithm. However, note from (12) that the bound in (11) is only valid whenever each R_u is upper bounded by $\frac{1}{N}\sum_{u=1}^{U} R_u$. Therefore, the optimization problem to solve is given by

$$\max_{\{Q_{u,m}\}_{u,m}} \sum\nolimits_{u=1}^{U} R_u \tag{13}$$

$$\text{Subject to: } R_u \le \frac{1}{N}\sum\nolimits_{k=1}^{U} R_k \quad \forall\ u \tag{14}$$

$$\sum\nolimits_{u=1}^{U}\sum\nolimits_{m=1}^{F} Q_{u,m} \le FP, \tag{15}$$

where[1]

$$R_u = \frac{B}{F}\sum_{m=1}^{F}\log\left(1+\frac{Q_{u,m} r_u^{-\alpha}\eta_{u,m}}{\sigma^2}\right). \tag{16}$$

Using the Lagrange multiplier method, we can define

$$\mathcal{L}(Q_{1,1},Q_{1,2},\dots,Q_{U,F-1},Q_{U,F})$$
$$=\sum_u R_u-\sum_u\left(\lambda_u\left(NR_u-\sum_k R_k\right)\right)-\rho\left(\sum_{u,m}Q_{u,m}-FP\right)$$
$$=\sum_u\left(1-N\lambda_u+\sum_k\lambda_k\right)R_u-\rho\left(\sum_u\sum_m Q_{u,m}-P\right).$$

We know that the optimum solution should satisfy

$$\frac{\partial\mathcal{L}}{\partial Q_{u,m}}=\left(1-N\lambda_u^\star+\sum_{k=1}^{U}\lambda_k^\star\right)\frac{1}{\frac{\sigma^2}{\eta_{u,m}}+Q_{u,m}^\star}-\rho^\star=0. \tag{17}$$

This optimization differs from the traditional water-filling since some of the users are saturated by the constraint (14). We divide the users to two groups, namely, the saturated users $\mathcal{U}_\text{S}$, for which (14) is satisfied with equality, and the unsaturated users $\mathcal{U}_\text{U}$, for which (14) is satisfied with inequality. We note that for the unsaturated users, the rate upper-bound constraint is loose and hence the optimum multipliers in the Lagrangian satisfy $\lambda_u^\star=0$ for $u\in\mathcal{U}_\text{U}$. Plugging this into (17), we get

$$\left(1-N\lambda_u^\star+\sum_{k=1}^{h}\lambda_k^\star\right)\frac{1}{\frac{\sigma^2}{\eta_{u,m}}+Q_{u,m}^\star}-\rho^\star=0 \quad u\in\mathcal{U}_\text{S},$$
$$\left(1+\sum_{k=1}^{h}\lambda_k^\star\right)\frac{1}{\frac{\sigma^2}{\eta_{u,m}}+Q_{u,m}^\star}-\rho^\star=0 \quad u\in\mathcal{U}_\text{U}.$$

For a given water level, $\nu=\left(1+\sum_{k=1}^{h}\lambda_k^\star\right)\frac{1}{\rho^\star}$, the power allocated to the unsaturated users is immediately given by

$$Q_{u,m}^\star=\left(\nu-\frac{\sigma^2}{\eta_{u,m}}\right)_+, \qquad u\in\mathcal{U}_\text{U}. \tag{18}$$

Plugging (18) in (16), we can find the rate for these users.

While the optimum power for unsaturated users can be simply determined by the water-filling algorithm, for saturated users the optimum power is determined by an additional step for each user, which determine the minimal power level to achieve the saturation rate. The saturation rate is given by:

$$R_\text{S}=\min_{0\le|\mathcal{U}_\text{U}|\le N-1}\frac{\sum_{u\in\mathcal{U}_\text{U}}\tilde{R}_u}{N-|\mathcal{U}_\text{U}|},$$

where $\tilde{R}_u$ be the pre-saturation user rates (i.e., the rates for the given water level, ν, but with $\lambda_u^\star=0$). Thus, given a water level, ν, we can find the rates and power allocation for each bin of each user. Now we need to find the water level that will achieve the actual power constraint. For that, we note that the resulting power from a given water level is monotonically increasing with the water level. Thus, the optimal water level can be found by a simple one-dimensional search.

In the following we further elaborate on the power allocation by revisiting the example we discussed before.

A. Continue with the Illustrative Example

Consider the example of Section III-A, and assume a total power budget of $P=159/64$ Watts. We also assume

$$\frac{\eta_1}{\sigma^2}=64,\qquad \frac{\eta_2}{\sigma^2}=\frac{\eta_3}{\sigma^2}=4,\qquad \frac{\eta_4}{\sigma^2}=2.$$

Using the standard water-filling algorithm we get water level of $\rho=5/8$, and power allocation of

$$\hat{P}_1=55/64,\qquad \hat{P}_2=\hat{P}_2=5/8,\qquad \hat{P}_4=3/8,$$

which result in rates

$$\hat{R}_1=\log(56),\quad \hat{R}_2=\hat{R}_2=\log(7/2),\quad \hat{R}_4=\log(7/4).$$

However, these rates violate the constraint in (14), as $\hat{R}_1=5.81>5.11=\sum_{u=1}^{4}\hat{R}_u/2$. Solving the constrained power allocation problem using the proposed algorithm, we obtain

$$P_1^\star=31/16,\qquad P_2^\star=P_3^\star=3/4,\qquad P_4^\star=1/2,$$

and

$$R_1^\star=5,\qquad R_2^\star=R_3^\star=2,\qquad R_4^\star=1.$$

V. Numerical Results

In this section we present numerical results that better illustrate the gain obtained by cache-aided communication with optimal power allocation. In all simulations we consider a BS with $N=4$ antennas and $N+M$ users. All files are of 1MB ($8\cdot10^6$ bits). The transmission bandwidth is set to 1MHz and is divided into $B=100$ frequency bins. The power is set such that the average SNR for users at 1Km distance is 0dB.

We first consider a synthetic scenario in which $N+M-1$ users are located at a distance of 1Km from the BS, while User 1 is located at a varying distance. Fig. 2 depicts the total transmission time of the network, T.

The dashed lines represent a simple equal power allocation between all users and between all frequency bins. This scheme is obviously sub optimal. But we need to also note that it does not satisfy the rate constraint of (7). Thus, any power that is spent in achieving rates beyond the constraint is waisted. The solid lines depicts the performance using the optimal power allocation scheme described above.

[1] By letting $F\to\infty$ we can obtain the expected rate.

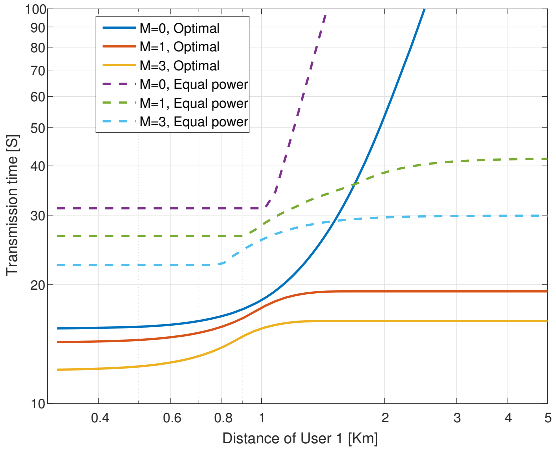

Fig. 2: Network transmission time as a function of the Distance of User 1, when all other users are located at a distance of 1Km.

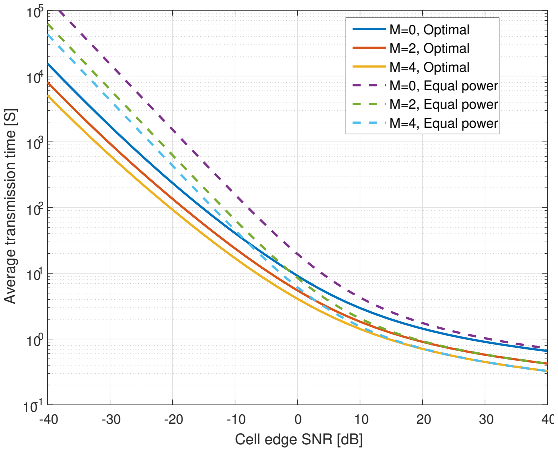

Fig. 3: Average network transmission time as a function of the cell edge SNR in a random network with a radius of 1Km.

In the case of no cache ($M = 0$), in the case of equal power allocation, the fairness constraint dictates that the total transmission time is the time that is needed by the weakest user. Adding cache brings a significant improvement in performance for both schemes. The optimal cache placement gives an alternative approach to balance the users, and hence even with equal power allocation we see significant gain. The transmission time is obviously much shorter when we combine optimal cache and power allocation.

To give a complete system view, we also performed random network simulations. The presented results are based on Monte Carlo simulation with 1000 network realizations, each with a random positioning of the $U = M + N$ users independently and uniformly over a circular area of radius 1Km.

Fig. 3 depicts the average transmission time in the network as a function of the average SNR at the cell edge (i.e., as the total transmission power changes). As expected, at high SNR, the equal power is close to optimal, and all rates become quite close. In such case, the DoF characterize the performance and the time is proportional to $1/(M + N)$ (i.e., a time reduction of $\frac{2}{3}$ and $\frac{1}{2}$ for $M = 2$ and $M = 4$, respectively). For low signal to ratio we see that the optimal power allocation has a significant gain.

VI. Conclusion

In this paper we focused on the joint cache and power allocation in cache-aided communication. We considered a cellular network with multiple antennas at the base station and asymmetric links, and focused on the special case where the number of users equals the number of BS antennas plus the number of copies of the dataset distributed across all users.

Interestingly, it is shown that the joint optimization of cache usage and power allocation yields a gain for MISO coded-caching, which goes beyond the throughput increase. In particular, it is shown that the cache is used to balance the users such that fairness and throughput are no longer contradicting. More specifically, in this case, strict fairness (all users can decode files of the same size) is achieved jointly with maximizing the network throughput.

References

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

[2] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, “The exact rate-memory tradeoff for caching with uncoded prefetching,” *arXiv preprint arXiv:1609.07817*, 2016.

[3] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, “Multi-antenna coded caching,” *arXiv preprint arXiv:1701.02979*, 2017.

[4] M. Gregori, J. Gómez-Vilardebò, J. Matamoros, and D. Gündüz, “Joint transmission and caching policy design for energy minimization in the wireless backhaul link,” in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 1004–1008.

[5] A. Ghorbel, M. Kobayashi, and S. Yang, “Content delivery in erasure broadcast channels with cache and feedback,” *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6407–6422, 2016.

[6] S. Yang, K.-H. Ngo, and M. Kobayashi, “Content delivery with coded caching and massive mimo in 5g,” in *Turbo Codes and Iterative Information Processing (ISTC), 2016 9th International Symposium on*. IEEE, 2016, pp. 370–374.

[7] 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.

[8] A. M. Ibrahim, A. A. Zewail, and A. Yener, “Optimization of heterogeneous caching systems with rate limited links,” in *Communications (ICC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1–6.

[9] X. Peng, J. Zhang, S. Song, and K. B. Letaief, “Cache size allocation in backhaul limited wireless networks,” in *IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–6.

[10] T. X. Vu, S. Chatzinotas, and B. Ottersten, “Coded caching and storage planning in heterogeneous networks,” in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2017, pp. 1–6.

[11] J. Hachem, N. Karamchandani, and S. N. Diggavi, “Content caching and delivery over heterogeneous wireless networks,” in *IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 756–764.

[12] S. S. Bidokhti, M. Wigger, and A. Yener, “Benefits of cache assignment on degraded broadcast channels,” in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 1222–1226.

[13] A. Ghorbel, K.-H. Ngo, R. Combes, M. Kobayashi, and S. Yang, “Opportunistic content delivery in fading broadcast channels,” *arXiv preprint arXiv:1702.02179*, 2017.

[14] M. M. Amiri and D. Gündüz, “Decentralized caching and coded delivery over gaussian broadcast channels,” in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 2785–2789.

[15] I. Bergel and S. Mohajer, “Cache aided communications with multiple antennas at finite snr,” 2017, submitted to JSAC.