Cache-Aided Communications With Multiple Antennas at Finite SNR

Itsik Bergel, *Senior Member, IEEE*, and Soheil Mohajer, *Member, IEEE*

***Abstract*— We study the problem of cache-aided communication for cellular networks with multi-user and multiple antennas at finite signal-to-noise ratio. Users are assumed to have non-symmetric links, modeled by wideband fading channels. We show that the problem can be formulated as a linear program, whose solution provides a joint cache allocation along with pre-fetching and fetching schemes that minimize the duration of the communication in the delivery phase. The suggested scheme uses zero-forcing and cached interference subtraction, and hence, allows each user to be served at the rate of its own channel. Thus, this scheme is better than the previously published schemes that are compromised by the poorest user in the communication group. We also consider a special case of the parameters for which we can derive a closed form solution and formulate the optimal power, rate, and cache optimization. This special case shows that the gain of MIMO coded caching goes beyond the throughput. In particular, it is shown that in this case, the cache is used to balance the users such that fairness and throughput are no longer contradicting. More specifically, in this case, strict fairness is achieved jointly with maximizing the network throughput.**

***Index Terms*— Cache-aided communication, MIMO, finite SNR regime, MIMO, cache and power allocation, linear optimization, zero-forcing.**

I. Introduction

NETWORK traffic has rapidly increased, over both wired and wireless networks, in recent years. This overwhelming growth is mostly due to the demands for broadband data. In particular, video delivery accounts for a major growth of traffic on both mobile [1] and wireline networks [2]. Two unique characteristics of of video contents are (i) popular files are repeatedly requested by multiple users; and (ii) unlike general web usage, video request has a prime time. These unique properties provide an opportunity for storing the data at local caches during the off-peak hours of the network, and serve a request at the peak hours [3].

In a pioneering work Maddah-Ali and Niesen [4], showed that caching gain is not limited to the local cache size at individual users. More importantly, caching a packet at User 1, even if it is only requested by another User 2, provides an opportunity for *multicasting combined packets*, which can simultaneously serve both Users 1 and 2. It is shown that this scheme offers a global gain which scales with the aggregate size of the caches distributed across all users in the network.

This scheme was further generalized to multiple transmit antennas [5], where they showed 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 complete dataset stored across over all users. This is a significant gain, in contrast to only N DoF, achievable with N antennas and no caching. This potential gain is substantial, even in spite of the current trend of massive MIMO [6], [7], which calls for the use of antenna arrays with many elements: especially due to the cost of antennas arrays to be deployed. In contrast, use of a small cache at each mobile comes with very low cost, and these memories easily scale up to a total size that can offer significant gains.

Nevertheless, the existing works in this area only focus on isolated scenarios, or limited to DoF characterization (i.e., asymptotically high SNR). Thus, there is a big gap to cover before we can understand and fully 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 optimal use of caching in such a multi-antenna scenario is still unknown.

In this work, we demonstrate the achievability of the caching gain in the finite SNR regime and present closed form expressions for the performance in a special case. These results demonstrate a fascinating phenomenon, in which the cache contributes both for throughput and fairness. Recalling that in general fairness in wireless networks comes at the price of reduced throughput, the proposed scheme brings a situation in which strict fairness is achieved through maximizing the total network throughput. In other words, the caching allows a natural balancing of the load between the users. Thus, caching brings two distinct advantages: it increases the network throughput and it balances between the data rate of the different users to improve fairness.

A short illustrative example that demonstrate the core of the proposed scheme is given in Section III-A.

A. Related Works

Coded caching [4] is a novel data delivery technique to exploit the aggregate cache in the network rather than individual memory available at each user. In general, we have a network with U users and a set of F files at the server, all of the same (normalized) length. Each user is equipped with a storage memory to store a fraction of the packets of each file during the *placement phase*. Cache placement occurs prior

Manuscript received December 11, 2017; revised May 25, 2018; accepted May 25, 2018. Date of publication June 6, 2018; date of current version October 30, 2018. The work of S. Mohajer was supported by the National Science Foundation under Grant CCF-1749981. *(Corresponding author: Itsik Bergel.)*

I. Bergel is with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel (e-mail: itsik.bergel@biu.ac.il).

S. Mohajer is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: soheil@umn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2018.2844618

to users' requests, and hence it is performed regardless and independent of the requests. At the beginning of the delivery phase, each user requests for one file from the dataset, and the server broadcasts a message (a sequence of packets) to simultaneously serve all user requests. That is, each user u should be able to recover its desired file from the received message and its cached information. The ultimate goal is to minimize the duration of time required to serve all users.

A *global caching gain* can be realized when a single transmission can serve multiple users: A packet requested by user u and not cached at his local memory can be *combined* with any other packet cached at the user, since it allows the user to null the *interference* using its local memory. Such a combined packet is simultaneously useful for some other user u', if it contains a packet requested by u', and all other interfering components are cached in u'.

Several interesting bounds have been proposed to fully characterize the rate-memory tradeoff of coded caching [8]–[12]. Under ideal assumptions and uncoded prefetching, it is shown that the proposed scheme of [4] is information-theoretically optimal for some range of parameters [13]. Optimality of (a slightly modified version of) this scheme is proved in [14] for arbitrary parameters and for both average and worst-case demand scenarios.

In general, the caching gain can be improved by allowing coded pre-fetching (referring to jointly coding across files to be placed at usersâĽ™ cache), at the price of complexity of the system [15]–[18]. In spite of developing better achievability schemes, and several efforts in tightening the outer bound of the rate-memory tradeoff for caching with general placement [10], [19], [20], the problem is still not fully solved. One of the main advantages of uncoded pre-fetching is a rather simple handling of practically-relevant asynchronous demands, without increasing the communication rates [21], and hence we only focus on uncoded placement in this work.

Recently, the attention of the community has been shifted towards the practical aspects of coded caching, and their adoption in wireless networks. In particular, [22]–[27] study coded caching in wireless networks in the presence of fading and/or erasure channels. Coded caching in wireless networks with multiple antennas at transmitters and/or receivers is considered in [5], [24], [28], and [29]. In particular, a homogeneous (with statistically identical channel gains) MISO network is considered in [29], where a mixed communication scheme is proposed to combine spatial multiplexing and multicasting, and improve the gain as the number users grows.

Employment of coded caching in wireless networks, and in particular in cellular networks, requires addressing several practical issues. In a realistic system, each user has a channel with different statistics and capacity. Cache allocation should be optimized depending on network traffic, user's channel quality, user's available storage, and other network characteristics. Coded caching for heterogeneous networks with different channels and rates for users (in the delivery phase) is studied in [30] for networks with single transmit and receive antennas. In [30], each packet transmission is subject to the rate of the weakest user, among those supposed to decode the packet.

One of the fundamental distinctions of our work is the exploitation of spatial diversity for the delivery phase. In particular, joint coding of the packets can be performed over-the-air, instead of at the transmitter: The transmitter sends different packets along various spatial directions. Each end-user will receive a combination of the transmit packets. The interfering packets are either nulled over the air by zero-forcing, or suppressed at the receiver using the cache content.

Hence, the rate of each packet is only limited by the channel capacity of the intended user. A rather similar phenomena is observed in the single antenna case, by using multiple nested codebooks [22], [31], [32]. In MIMO setting, however, this can be done naturally, since each message is sent along a different spatial direction, and users can suppress the effect of the undesired but cached messages from the received signal, even before the decoding process starts. This is an important characteristics of MIMO caching systems, which is further elaborated below.

The main contribution of this paper is the design of a cache aided communication scheme that serves each user at its own rate. This is done by using spatial multiplexing (instead of multicasting) and hence does not require each packet to be sent at the rate of the weakest user. The proposed system is DoF optimal, but gives significant advantage over previous methods were the rate of the users are different (typically at low and medium signal to noise ratio). We also derive a closed form solution and formulate the optimal power, rate and cache allocation for a special case of the parameters.

Our results indicate a significant improvement in the system throughput due to jointly optimizing cache, power and rate allocation. This is in contrast to the result of [33], where it is shown that a separate design of the caching and delivery is order-wise optimal. However, it is worth noting that while we have *total cache size constraint*, the setting in [33] associated a fixed and uniform cache size to each user, and hence its result does not directly apply to our setting.

The remainder of the paper is organized as follows. In Section II we present the system model. In Section III we formulate that caching optimization problem as a linear program (LP). A closed form solution for the problem for some special range of parameters is presented in Section IV, followed by some numerical results that illustrate the gain offered by caching and our proposed resource allocation method in Section V. Finally, we finish the paper by some concluding remarks in Section VI.

II. System Model

A. Network and Channel Model

We consider a single cell network with one base station (BS) which is serving U users. The BS has N_{T} antennas and each user is equipped with N_{R} antennas. We assume a strict fairness setup, in which each user requests exactly one file, and all files are of the same size (i.e., all users require exactly the same amount of data). We further assume a wideband communication scheme, in which the bandwidth is divided into B small frequency bins. Symbols are transmitted at the rate of R_{S} symbols per second, where at each symbol time

one symbol is modulated over each frequency bin without inter symbol interference (e.g., OFDM). Thus, the transmission bandwidth is approximately $B \cdot R_{\mathrm{S}}$.

Considering the time duration of a single symbol, the received sample after matched filtering for the m-th frequency bin at i-th user is described by an $N_{\mathrm{R}} \times 1$ vector, given by

$$\mathbf{y}_{i,m} = \mathbf{H}_{i,m}\mathbf{x}_m + \mathbf{w}_{i,m}, \tag{1}$$

where $\mathbf{H}_{i,m} \in \mathbb{C}^{N_{\mathrm{R}} \times N_{\mathrm{T}}}$ is the channel matrix between the BS and the i-th user in frequency bin m, which contains the gain from each BS antenna to each antenna of user i, $\mathbf{x}_m$ is the transmitted vector at this frequency bin and $\mathbf{w}_{i,m} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ is the additive complex white Gaussian noise.

We assume a very limited movement for the users during the transmission block. Thus, r_i, the distance between the i-th user to the BS, does not change. On the other hand, due to small movements of the users, and movements of other objects in the area, each link between two antennas experiences fading. Thus, the channel matrix 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 $\mathbf{G}_{i,m}$ is a random matrix that represents the fading. We consider a rich scattering environment, and hence each link experiences an independent Rayleigh fading. In mathematical terms, we assume that each element of $\mathbf{G}_{i,m}$ is a proper complex normal random variable, with zero mean and unit variance, and that all elements of the matrices $\mathbf{G}_{i,m}$ for $i = 1, \ldots, U$ are statistically independent. Note that we do not assume any specific model for the frequency dependence of the fading. Yet, we will later assume that the bandwidth is large enough, so that the aggregate rate over all frequency bins mimics the expected rate.

B. Transmission Scheme

The BS simultaneously transmits different messages to different users. Transmission of some of the requested messages can be ignored if the messages are already stored in the users' cache, as will be detailed later. For other messages, the BS needs to make sure that the transmission of an undesired message will not interfere with the reception of the desired user at each active 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. This approach is commonly termed Zero Forcing (ZF) precoding or more specifically, block-diagonalization [34]. To allow this scheme, we assume that the number of transmit and receive antennas satisfy $N_{\mathrm{T}} = N \cdot N_{\mathrm{R}}$.

While the block-diagonalization scheme adopted in this work is suboptimal, one should note that it has many merits. On one hand, this scheme is known to asymptotically achieve the optimal DoF in a multi-user MIMO scenario [34]. On the other hand, it requires a low implementation complexity, and hence is quite popular for practical implementations. Specifically to our work, it is convenient as it results in user rates that are independent of the other users rate and channel (as will be shown below).

To apply the block diagonalization constraint, we use a projection matrix, $\mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}}$, that projects to the null-space of the channel matrices of $N-1$ selected users, and $\mathcal{Z}_{i,m}$ is the set of users that should not be disturbed by the transmission to user i, that is, a transmit message to user i over frequency bin m should be zero-forced at users in $\mathcal{Z}_{i,m}$. Thus, the effective channel of user i at frequency bin m is $\mathbf{H}_{i,m}\mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}}$, and its achievable rate is:

$$\tilde{R}_{i,m} = R_{\mathrm{S}} \log_2 \left| \mathbf{I} + \frac{P_i}{\sigma^2} p_{i,m} \mathbf{H}_{i,m} \mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}} \mathbf{H}^H_{i,m} \right|, \tag{3}$$

where $|\cdot|$ denotes matrix determinant, 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, i.e, $p_{i,m}P_i$ is the effective power allocated to user i in frequency bin m. We will assume throughout that the intra-user power allocation is normalized to 1 ($\mathbb{E}[p_{i,m}] = 1 \ \forall i$), and the inter-user power allocation is subject to a sum-power constraint, $\sum_i P_i \leq P$.

C. Performance Evaluation

We assume that each user can decode its desired message without interference from other messages that were simultaneously transmitted by the BS (i.e., each interfering message is either zero forced by the BS or subtracted using the cache available at the receiver). Thus, the achievable rate for user i is given by:

$$\tilde{R}_i = \sum_m \tilde{R}_{i,m} = \sum_m R_{\mathrm{S}} \log_2 \left| \mathbf{I} + \frac{P_i}{\sigma^2} p_{i,m} \mathbf{H}_{i,m} \mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}} \mathbf{H}^H_{i,m} \right|.$$

Assuming that the bandwidth is *large enough*, it will contain enough fading variations so that we can apply the law of large numbers. Let denote by B the number of frequency bins. Thus, for sufficiently large B the user rate will converge to its expectation:

$$\frac{\tilde{R}_i}{B} \to \frac{R_i}{B} = R_{\mathrm{S}} \mathbb{E}\left[\log_2 \left| \mathbf{I} + \frac{P_i}{\sigma^2} p_{i,m} \mathbf{H}_{i,m} \mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}} \mathbf{H}^H_{i,m} \right| \right]. \tag{4}$$

Substituting (2) into (4), we have:

$$R_i = B R_{\mathrm{S}} \mathbb{E}\left[\log_2 \left| \mathbf{I} + \frac{P_i r_i^{-\alpha}}{\sigma^2} p_{i,m} \mathbf{G}_{i,m} \mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}} \mathbf{G}^H_{i,m} \right| \right]. \tag{5}$$

For a fixed user i with given P_i and r_i, the random quantities $p_{i,m}$, $\mathbf{G}_{i,m}$ and $\mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}}$ only depend on the channel fading (i.e., matrices $\{\mathbf{G}_{i,m}\}$). Thus, the expectation in (5) (which is taken with respect to the fading) depends only on P_i and r_i. Hence, we conclude that the user rate depends only on its distance to the BS, r_i, and its allocated power, P_i. In this setup, it is convenient to characterize each user solely by its achievable rate, R_i.

As an example, in the single receive antenna case ($N_{\mathrm{R}} = 1$), the product $\mathbf{G}_{i,m}\mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}}\mathbf{G}^H_{i,m}$ is a rank-1 matrix (indeed it is an scalar), and its single eigenvalue (denoted as $\rho_{i,m}$) has a standard exponential distribution. If we also assume constant intra-user power allocation ($p_{i,m} = 1$), the user rate will be

$$\begin{aligned} R_i &= B \cdot R_{\mathrm{S}} \cdot \mathbb{E}\left[\log_2\left(1 + \eta_i \rho_{i,m}\right)\right] \\ &= -B R_{\mathrm{S}} \log_2 e \cdot e^{1/\eta_i} \cdot \mathrm{Ei}(-1/\eta_i), \end{aligned} \tag{6}$$

where $\eta_i = \frac{P_i r_i^{-\alpha}}{\sigma^2}$ is the average SNR and $\mathrm{Ei}(\cdot)$ is the exponential integral function, defined as $\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t} dt$.

D. Caching

A cache-aided communication scheme includes two phases, namely, placement phase and delivery phase. There is a database of F files, each of a unit length, available at the BS, and each user is interested in one of the files. Each user has an allocated cache, to pre-store some part of the database. During the placement phase, the users' caches are filled with messages (packets) from the database, while the users' requests are not yet revealed. After the placement phase, upon revealing users' demands, the BS transmits a proper set of packets in order to serve all the users with their desired files. The placement phase occurs in the off-peak time of the network, in order to improve the communication in the peak-time. Even though our analysis can be applied on general demand profile, in this work we consider the worst case demand scenario, in which users request distinct files (and consequently, we assume $F \geq U$).

The BS can decide on the best allocation of cache to users, subject to a total cache constraint of $M \cdot F$ units distributed over all users. This optimization allows the BS to place larger cache at users with lower rates (poor channel conditions) and hence reduce the total transmission time of the BS.

We assume that each user is equally likely to request any file. Thus, without loss of generality, we can simplify the problem by assuming that each user will store similar parts of all files, and hence, the cache contents of the users is invariant under a permutation of the files. Thus, the caching problem reduces to finding the optimal cache placement and the optimal sequence of transmissions that will deliver the desired files to all users in minimal time.

Note that the cache allocation problem does not depend on many of the systems parameters described above. In fact, it turns out that if the number of files is not less than the number of users ($F \geq U$), the optimum solution for the caching problem only depends on the number of users U, the spatial multiplexing dimension N (the number of users that can be simultaneously served with no interference using only the selected MIMO scheme), the number of copies of the database that are distributed across users' cache M, and the communication rates supported by the channel $\{R_i\}_i$.

III. Caching Optimization

A cache aided communication scheme needs to specify which part of each file to be pre-fetched at each user (during the placement phase), and afterwards, given the user requests, what is the transmission scheme that can satisfy the requests of all users (during the delivery phase), i.e., what parts of what files should be jointly transmitted at each stage so that all users will be able to decode all their desired packets. In Subsection III-B we show that this problem can be formulated as a linear optimization problem, and hence can be solved efficiently using linear programming methods. Before that, we give a simple example that illustrates the operation of a valid transmission scheme.

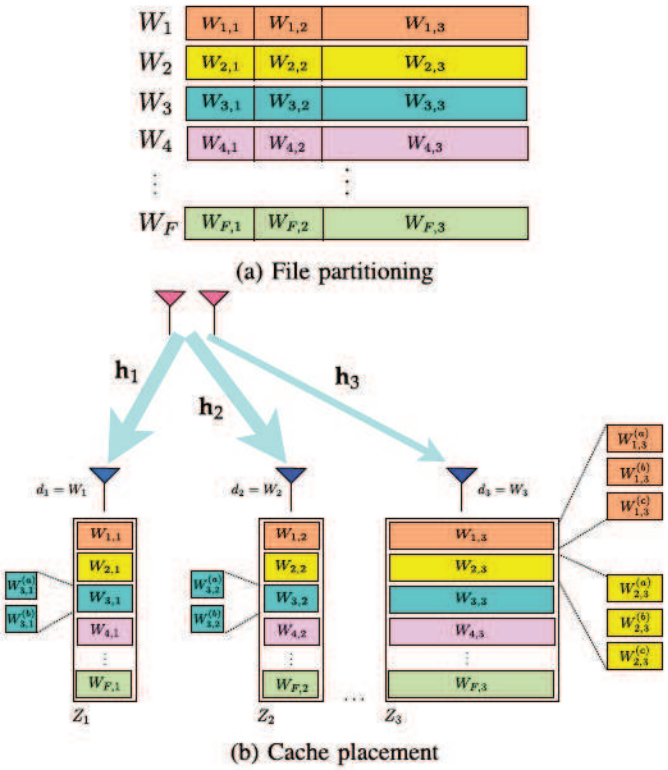

Fig. 1. Cache optimization based on user channel quality.

A. A Simple Example

Consider a MISO broadcast channel with $N_{\mathrm{T}} = 2$ transmit antennas and $U = 3$ users with single antenna ($N_{\mathrm{R}} = 1$), as shown in Fig. 1. We assume a total cache constraint, so that only one copy of each (packet of each) file can be pre-fetched among all the users, i.e., $M = 1$. We denote the fraction of files to be cached at user i by q_i, which implies $q_1 + q_2 + q_3 = M = 1$. Recall that the cache placement is invariant under file relabeling, and hence, q_i faction of each file in the dataset should be pre-fetched in user i. We denote the link capacity of user i by R_i. Assume Users 1 and 2 have good channels to support $R_1 = R_2 = 2$ and the third user is further away from the transmitter and can only decode at rate of $R_3 = 1$.

It turns out that the optimum cache allocation to compensate for the weakness of User 3 is $q_1 = q_2 = 1/5$ and $q_3 = 3/5$. The cache allocation is done by partitioning each file into 3 *sections*, namely, $W_{k,1}$, $W_{k,2}$, and $W_{k,3}$, which are stored at the cache of Users 1, 2, and 3 respectively, as shown in Fig. 1(b). The length of cached sub-files will be $|W_{k,1}| = |W_{k,2}| = 1/5$ and $|W_{k,3}| = 3/5$ (recall that the file lengths are normalized, and hence $1/5$ refers of $1/5$ of a the actual length of the files). Note that the cache placement is performed prior to the users request, and hence is identical for all files. In this example we assume that User 1 requested $d_1 = W_1$, User 2 requested $d_2 = W_2$ and User 3 requested $d_3 = W_3$.

The delivery phase includes broadcasting 4 messages. To formally present the broadcast messages, we need to uniformly divide some of the cached sections into smaller *segments* as $W_{3,1} = \left(W_{3,1}^{(a)}, W_{3,1}^{(b)}\right)$, $W_{3,2} = \left(W_{3,2}^{(a)}, W_{3,2}^{(b)}\right)$, $W_{1,3} = \left(W_{1,3}^{(a)}, W_{1,3}^{(b)}, W_{1,3}^{(c)}\right)$, and $W_{2,3} =$

$\left(W_{2,3}^{(a)}, W_{2,3}^{(b)}, W_{2,3}^{(c)}\right)$, to keep up with the capacity of the links to the users. Then, each section $W_{k,\cdot}$ and segment $W_{k,\cdot}^{(\cdot)}$ will be coded to a sequence $\mathrm{s}_{k,\cdot}$ and $\mathrm{s}_{k,\cdot}^{(\cdot)}$, respectively, using a channel code of rate R_k. Hence, the length of the resulting sequences will be $|\mathrm{s}_{1,2}| = |\mathrm{s}_{2,1}| = \left|\mathrm{s}_{k,\cdot}^{(\cdot)}\right| = \frac{1}{R_k}\left|W_{k,\cdot}^{(\cdot)}\right| = \frac{1}{10}$.

The sequences needed by the users for successfully decoding their requested files are

$$\begin{aligned}
&\text{User 1}: \ \mathrm{s}_{1,2}, \mathrm{s}_{1,3}^{(a)}, \mathrm{s}_{1,3}^{(b)}, \mathrm{s}_{1,3}^{(c)},\\
&\text{User 2}: \ \mathrm{s}_{2,1}, \mathrm{s}_{2,3}^{(a)}, \mathrm{s}_{2,3}^{(b)}, \mathrm{s}_{2,3}^{(c)},\\
&\text{User 3}: \ \mathrm{s}_{3,1}^{(a)}, \mathrm{s}_{3,1}^{(b)}, \mathrm{s}_{3,2}^{(a)}, \mathrm{s}_{3,2}^{(b)}.
\end{aligned}$$

All three users can be served by transmitting:

$$\begin{aligned}
\mathrm{x}(1) &= \mathrm{h}_2^{\perp}\mathrm{s}_{1,3}^{(a)} + \mathrm{h}_1^{\perp}\mathrm{s}_{2,3}^{(a)} + \mathrm{h}_2^{\perp}\mathrm{s}_{3,1}^{(a)},\\
\mathrm{x}(2) &= \mathrm{h}_3^{\perp}\mathrm{s}_{1,2} + \mathrm{h}_1^{\perp}\mathrm{s}_{2,3}^{(b)} + \mathrm{h}_2^{\perp}\mathrm{s}_{3,1}^{(b)},\\
\mathrm{x}(3) &= \mathrm{h}_2^{\perp}\mathrm{s}_{1,3}^{(b)} + \mathrm{h}_3^{\perp}\mathrm{s}_{2,1} + \mathrm{h}_1^{\perp}\mathrm{s}_{3,2}^{(a)},\\
\mathrm{x}(4) &= \mathrm{h}_2^{\perp}\mathrm{s}_{1,3}^{(c)} + \mathrm{h}_1^{\perp}\mathrm{s}_{2,3}^{(c)} + \mathrm{h}_1^{\perp}\mathrm{s}_{3,2}^{(b)},
\end{aligned} \tag{7}$$

in four time slots, where $\mathrm{x}(t)$ is beam-forming transmission vectors for the t-th time block, which takes $1/10$ time slots. The notation $\mathrm{h}_i^{\perp}\mathrm{s}_{k,j}$ indicates that all symbols of the codeword $\mathrm{s}_{k,j}$ are precoded over all frequency bins and several symbols, and each pre-coding vector at the m-th frequency bin is perpendicular to the i-th user channel, $\mathrm{h}_{i,m}$.

Let us consider file retrieval at User 1. For instance, in time block $t = 1$ and frequency bin m, User 1 receives $y_{1,m}(1) = \mathrm{h}_{1,m}\mathrm{x}_m(1) + w_m(1) = \mathrm{h}_{1,m}\mathrm{h}_{2,m}^{\perp}(\mathrm{s}_{1,3,m}^{(a)} + \mathrm{s}_{3,1,m}^{(a)}) + w_m(1)$. It removes $\mathrm{s}_{3,1,m}^{(a)}$ using its cache, and then uses the remaining signal to decode $W_{1,3}^{(a)}$. Similarly, each user can decode all the missing sections of its requested file.

Note that each transmission takes $1/10$ time slots, and hence the total transmission time is $4/10$, after which, all users have their requested files. A total of 3 files (each of unit length) are delivered to the users, where the network delivered a total of $4/5 + 4/5 + 2/5 = 2$ files and the remaining sections were already stored at the cache of requesting users. Thus, the throughput of the network is $2/0.4 = 5$. In contrast, in a similar setting with only single-antenna transmitter, the rate of each packet intended for a subset of users including User 3 should not exceed $R_3 = 1$. This shows that an optimized coded caching in MISO *offers more gain than just trading antennas vs. cache memory*.[1]

This example is further illustrated in Subsection III-B, using the terminology of an optimization problem (see Equation (11) and the preceding paragraph). □

B. Cache-Aided Communication as an Optimization Problem

We next derive a mathematical framework that can describe a cache-aided communication scheme, and show that it can be formulated as a linear programming problem. We focus only on *efficient* transmission schemes, where we define an 'efficient' transmission as one that exploits all degrees of freedom of the channel. In the setup at hand, an 'efficient' communication must serve $M+N$ users simultaneously at all times.[2] This is done by zero forcing each transmission to $N-1$ direction, and allowing M users to subtract the interference using their cache.

Thus, the content of each transmit message in the network must be stored by M users. In other words, each transmission is intended for a combination of $M+1$ users. Furthermore, as these users store the same parts of all files to allow 'efficient' transmissions, this specific user combination must use their cache for the transmission of specific file parts of every other user in the same group. To formulate that, we divide each file into L segments, where each segment is stored by M users. As we have a total of U users and segments are stored by M users, the maximal number of needed sections can be bounded by $L \leq \binom{U}{M}$. We enumerate these sections by $\ell = 1, 2, \ldots, L$, and describe them by the row vectors b_ℓ for $\ell = 1, 2, \ldots, L$, where $b_\ell(i) = 1$ if user i stores the ℓ-th section of each file, and $b_\ell(i) = 0$ otherwise.

The length of the ℓ-th section is denoted by u_ℓ. Thus, $0 \leq u_\ell \leq 1$ and $\sum_\ell u_\ell = 1$. Note that, the total fraction of each file cached at user i is given by $q_i = \sum_\ell b_\ell(i)u_\ell$, which implies

$$\sum_{i=1}^{U} q_i = \sum_{i=1}^{U}\sum_{\ell=1}^{L} b_\ell(i)u_\ell = \sum_{\ell=1}^{L}\left(u_\ell \sum_{i=1}^{U} b_\ell(i)\right) = \sum_{\ell=1}^{L} M u_\ell = M,$$

which guarantees that a total of M copies of the entire dataset is distributed among all users. Note that the vectors b_ℓ describe the different possibilities for file partitioning, and hence are known in advance (and depend only on M and U). The actual allocation is determined by the set of variables u_ℓ's which needs to be solved according to the available user rates.

The transmission in each time slot involves a combination of $M+N$ out of the U users, which can be simultaneously served. We will use an index c to label possible combinations where $c = 1, \ldots, C$ and $C = \binom{U}{M+N}$. Furthermore, each user combination can be active in several transmissions, each with different segments of the file transmitted to each user. The different transmissions for the same user combination (c) will be indexed by j.

Each of the $M+N$ users that are active in this time slot receives part of their requested file. We will use matrices $\mathbf{E}_j^c$ to describe the transmission scheme for a slot, where $(\mathbf{E}_j^c)_{\ell,i} = 1$ if the i-th user receives (part of) the ℓ-th file section at the j-th transmission of user combination c, and otherwise $(\mathbf{E}_j^c)_{\ell,i} = 0$. Thus, $1 \leq i \leq U$, $1 \leq \ell \leq L$ and $0 \leq c \leq C$. We will

[1] Note that many works (e.g., [5]) evaluate the rate based on the total delivered files (including the parts already cached at the users during the placement phase). In such terminology, the throughput of this network is $3/0.4 = 7.5$, as 3 files are delivered in $4/10$ time slots. We use the net throughput in our work in order to emphasize the relation to the physical rates, i.e., the network throughput is $R_1 + R_2 + R_3 = 5$.

[2] Using the notation of [5], where there are $\tilde{K}$ users, each with cache size of $\tilde{M}$, and a library of $\tilde{N}$ file, the union of the cache across users holds $\tilde{K}\tilde{M}/\tilde{N}$ copies of the entire data base (similar to our M). Using $\tilde{L}$ to denote the number of antennas in [5], and the total throughput definition, they showed that the per user DoF is $\frac{\tilde{L}+\tilde{K}\tilde{M}/\tilde{N}}{\tilde{K}(1-\tilde{M}/\tilde{N})}$. Replacing the notation, and also multiplying by $\tilde{K}$ for sum-DoF and multiplying by $(1-\tilde{M}/\tilde{N})$ to change from total throughput to net throughput (see footnote 1) we get a maximal sum-DoF of $N+M$. Thus, according to [5], any 'efficient' scheme will serve $M+N$ users at any time of transmission, and is hence DoF optimal.

next discuss the possible values of $\mathbf{E}_j^c$ and hence the maximal number of transmissions for any user combination.

To characterize the matrices $\mathbf{E}_j^c$, we note that each such matrix satisfies the following conditions:

(C1) Each element in the matrix is either zero or one ($(\mathbf{E}_j^c)_{\ell,i} \in \{0,1\}$).

(C2) Each user can receive only one segment at a time, and hence, there is at most one 1 in each column of $\mathbf{E}_j^c$ (i.e., $\sum_\ell (\mathbf{E}_j^c)_{\ell,i} \in \{0,1\}$).

(C3) In an 'efficient' transmission, at each time slot there are $N+M$ active users and hence, each $\mathbf{E}_j^c$ matrix contains exactly $M+N$ ones ($\sum_\ell \sum_i (\mathbf{E}_j^c)_{\ell,i} = N+M$).

(C4) As a user does not need a file segment that is already stored in its cache, we must have $(\mathbf{E}_j^c)_{\ell,i} = 0$ for any ℓ and i such that $b_\ell(i) = 1$ (or alternatively stated: $(\mathbf{E}_j^c)_{\ell,i} b_\ell(i) = 0$).

(C5) Only the users that belong to the user combination c will participate in the reception. Thus, all cache storage indicated by the matrix $\mathbf{E}_j^c$ must be of active users. In other words, if user i is not active in the matrix $\mathbf{E}_j^c$ (that is if $\sum_\ell (\mathbf{E}_j^c)_{\ell,i} = 0$), then it is also not used for cache storage ($\sum_\ell b_\ell(i)(\mathbf{E}_j^c)_{\ell,i} = 0$).

Thus, for a specific combination of $N+M$ out of U users, each matrix $\mathbf{E}_j^c$ contains $M+N$ ones in $M+N$ columns associated to the active users. Each one can be selected independently in its column from the allowed locations (where each column represents a user). In order to count the number of allowed locations of a one in a specific column, we note that (C4) requires the vector b_ℓ that corresponds to the row with the one must be zero for this user. Thus we need to count the number of vectors b_ℓ that has zero for this user. But, (C5) further limits the allowed locations as it requires that all relevant vectors, b_ℓ, must have their M ones chosen only from the $M+N$ active users. Thus, we need to count the number of vectors that have M ones out of $N+M-1$ users (the users that are active, excluding the considered user, that must be zero). Hence, there are a total of $\binom{M+N-1}{M}$ choices for the location of one in each column.

As the choices of the location of one in each column are independent, and there are $N+M$ columns of active users, in each $\mathbf{E}_j^c$ matrix, we have a total of

$$J = \binom{M+N-1}{M}^{(M+N)} \tag{9}$$

possible matrices for each users combination (i.e., the range of j is given by $1 \le j \le J$). Denoting by T_j^c the duration of time required to transmit to a user combination c in mode j, the total transmission time is given by:

$$T = \sum_{j,c} T_j^c. \tag{10}$$

Revisiting the Example: To demonstrate this formulation, consider the example of Section III-A. As $M=1$ and $U=3$ there only $L = \binom{U}{M} = 3$ file sections, and three vectors that describe their storage at the different users: $b_1 = [1,0,0]$, $b_2 = [0,1,0]$ and $b_3 = [0,0,1]$. The cache placement solution tells us the size of the segments are $u_1 = u_2 = 0.2$ and $u_3 = 0.6$. For the transmission scheme, we note that this case has only $C = \binom{3}{1+2} = 1$ user combination, and a total of $J = \binom{1+2-1}{1}^{(1+2)} = 8$ transmission schemes. Out of these, the obtained solution uses only 4 schemes:

$$\mathbf{E}_1^1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \quad \mathbf{E}_2^1 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$
$$\mathbf{E}_3^1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{E}_4^1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \tag{11}$$

and the transmission time of each mode is $T_1^1 = T_2^1 = T_3^1 = T_4^1 = 0.1$, implying a total transmission time of $T = 0.4$. □

Recall that transmission to User i is done at rate R_i. In order to serve User i we have to deliver all non-cached sections of the requested file, that is, ℓ's with $b_\ell(i) = 0$. Noting that the size of the ℓ section is u_ℓ, we have

$$\sum_{j,c} T_j^c (\mathbf{E}_j^c)_{\ell,i} = \frac{u_\ell}{R_i}. \tag{12}$$

for each User i and section ℓ such that $b_\ell(i) = 0$. Noting that each vector b_ℓ has $U-M$ zero elements, (12) yield in a set of $L(U-M)$ constraints.

Thus, the problem can be formulated as a linear programming minimization:

$$\begin{aligned} \min_{T_j^c, u_\ell} \ & \sum_{j,c} T_j^c \\ \text{Subject to} \ : & \sum_\ell u_\ell = 1 \\ & \sum_{j,c} T_j^c (\mathbf{E}_j^c)_{\ell,i} = \frac{u_\ell}{R_i} \quad \forall \ell, \ i : b_\ell(i) = 0 \\ & T_j^c \ge 0, \quad u_\ell \ge 0 \ \forall c, j, \ell \end{aligned} \tag{13}$$

The formulation of a linear programming problem allows us to find an optimal scheme using efficient algorithms. The solution of this problem gives both the details of the cache allocation and placement to all users (through the variables u_ℓ and equation $q_\ell = \sum_\ell b_\ell(i) u_\ell$), and the sequence of transmission that can deliver the desired files to all requesting users.

Note that the optimization problem above can easily be adjusted for the case of per user cache size constraint, by adding the constraint: $\sum_\ell b_\ell(i) u_\ell \le u_{\max}$, where $u_{\max}$ is the maximum cache size. Yet, in this work we do not pursue this approach, and focus only on the global cache size constraint, as described in Section II.

The number of variables (u_ℓ's and T_j^c's) in this problem is

$$L + C \cdot J = \binom{U}{M} + \binom{U}{M+N} \cdot \binom{M+N-1}{M}^{(M+N)}.$$

This number grows polynomially with the number of users U, but exponentially with $M+N$. Thus, the suggested approach is practical for large network as long as the number of DoF $(M+N)$ is not large. Further research is necessary to optimize large networks with large number of antennas or large cache.

On the good side, the number of equality constraints of the linear program is only $L(U-M)+1$. Thus, an optimal solution

includes at most $L(U-M)+1$ non-zero variables [35]. This means that the total number of file sections and transmission modes that are needed for the actual implementation is quite small and limited to $\binom{U}{M}\cdot(U-M)+1$.

As an example, consider a problem with $U=12$ users with (normalized) rates of $R_i=i$ for $i=1,\ldots,12$. Assume that the BS has $N_{\mathrm{T}}=2$ antennas, and $M=2$ copies of the dataset are distributed among all the users. The resulting linear programming problem has 40161 variables and 661 constraints. The actual solution divide the files to only 55 sections, and uses 421 transmission modes (i.e., the optimal solution resulted in 39674 transmission modes that were assigned a zero duration of time).

The total cache allocated (q_i) for each user in this scheme, given by $\sum_\ell u_\ell\cdot(b_\ell)_i$, is (sorted from the user with lowest rate to the user with highest rate): .63, .44, .29, .20, .14, .10, .07, .05, .03, .03, .02, 0. The total transmission time for the whole transmission is $T=0.51$.

As a comparison, a standard algorithm (e.g., mimicking [4] for multiple antenna case) that does not account for the different rates, will need to adjust each transmission to the active user with lowest rate. Such algorithm will need more than $T=1.22$ to complete all transmissions. This is more than twice slower then the proposed algorithm. As another comparison, using this optimal allocation but with $N=1$ BS antenna instead of 2 requires $T=0.73$ which is only 40% worse than the $N=2$ case, and still much better than the standard method with even 2 antennas.

IV. The Special Case of $U=M+N$

While the linear programming approach allows an efficient optimization of the cache aided communication scheme, it is hard to draw insights from it on the properties of the optimal solution. To get some insights, in the next section we analyze the special case of $U=M+N$.

A. Cache Allocation

In this special case, all users are active throughout all the transmissions. This allows for an analytical performance evaluation, as stated in the following theorem.

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

$$R_i\le\frac{1}{N}\sum_{u=1}^{U}R_u, \tag{14}$$

then the minimal time to serve all users is

$$T=\frac{N}{\sum_u R_u}. \tag{15}$$

Proof (Proof of Theorem 1): In the case that $U=M+N$ all 'efficient' transmissions must include 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

$$1-q_u=TR_u. \tag{16}$$

In addition, the total cache allocated across users satisfies:

$$\sum_{u=1}^{U}q_u=M. \tag{17}$$

If the optimization problem in (13) has a feasible solution, it must satisfy (16) and (17). Substituting (16) in (17) gives:

$$\sum_{u=1}^{U}(1-q_u)=(N+M)-\sum_{u=1}^{U}q_u=T\sum_{u=1}^{U}R_u,$$

which implies

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

However, this solution can be feasible only if the resulting q_i's are feasible. That is,

$$q_u=1-TR_u\ge 0,$$

which is equivalent to $R_u\le\frac{1}{T}=\frac{1}{N}\sum_{u=1}^{U}R_u$.

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. The cache placement in this case only needs to satisfy two simple requirements: 1) Exactly q_u of each file in the database should be stored at user u, and 2) The cache content at each user has no overlaps. 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, these 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.

B. Power Allocation

The result of Theorem 1 implies that a network with $U=N+M$ users can achieve a total throughput of[3] $\sum_{u=1}^{M+N}R_u$. This suggests that the well known water-filling algorithm will be appropriate for throughput optimization. However, the results above also include a condition that the maximal rate should not exceed $\frac{1}{N}\sum_{u=1}^{M+N}R_u$. An intuitive justification for this constraint is the following: A user with a rate that is higher than $\frac{1}{N}\sum_{u=1}^{M+N}R_u$ will be completely served before the other users. Hence, for the remaining transmission time, there are less than $N+M$ active users in the system, and we cannot fully exploit the available DoF of the network.

This result also represents a fascinating balancing mechanism that brings a natural balance between throughput and fairness. The issue of throughput maximization vs. user fairness has accompanied the field of wireless communication for decades. In most cases, enforcing fairness reduces the total throughput, and the typical working point is selected as a trade-off between the two.

In this work, the problem is stated with a strict fairness constraint: each user must receive the same amount of

[3]Taking into account the files stored in cache, the throughput is defined as $(U-M)/T$ where T is the time needed to complete transmission for all users.

data (1 file). Yet, due to the caching, the optimal performance is achieved by maximizing the total throughput. Thus, the caching allowed a natural balancing of the load between the users. Hence, caching brings two distinct advantages: (1) it increases the network throughput by allowing a simultaneous transmission of $N+M$ users, and, (2) it balances between the data rate of the different users to improve fairness. Note that the second property is obtained mostly by placing larger cache to poor users, which reduces their communication needs.

The maximum rate constraint represents the cases in which maximal throughput cannot be jointly achieved with the complete fairness. In such cases, the system needs to allocate more power to poor users in order to further increase their rate and achieve fairness. At the power allocation level, we can allocate the total power among the users to guarantee achievability of the maximum throughput. This can be formally stated as an individual optimization problem stated in (18).

$$\max_{P_1,\ldots,P_U} \sum_{u=1}^{U} R_u \qquad \text{Subject to: } R_k \le \tfrac{1}{N}\sum_{u=1}^{U} R_u \quad \forall\, k \tag{18}$$

The solution for this problem can be obtained by a small adjustment of the standard water-filling algorithm. For simplicity, the algorithm is described only for the case of single antenna per user. Let the power level be denoted by ρ. For convenience, we sort the users by their distance from the BS (such that $r_U \le r_{U-1} \ldots \le r_1$) and denote the effective channel gain by $\eta_{i,m} = \mathbf{G}_{i,m}\mathbf{P}^{\perp}_{\mathcal{Z}_{i,m}}\mathbf{G}^{H}_{i,m}$. The optimum water-filling power and rate allocations are given by:

$$P_i p_{i,m} = \left(\rho - \frac{\sigma^2}{r_i^{-\alpha}\eta_{i,m}}\right)_+,$$
$$R_i^{\mathrm{W}} = \mathbb{E}\left[\log_2\left(1 + \frac{P_i p_{i,m} r_i^{-\alpha}\eta_{i,m}}{\sigma^2}\right)\right], \tag{19}$$

where $x_+ = \max(x,0)$ is the positive part of x. If these rates do not satisfy the maximal rate constraint, we need to determine which users will meet the constraint with equality. Noting that these users will always be the users closest to the BS, we just need to determine the number of users for which the constraint will be active. Denoting the number of such users by h, these users will use the constraint rate, $R_{\max}(h)$. Thus, the maximal allowed rate will be

$$R_{\max}(h) = \frac{1}{N}\sum_{u=1}^{U-h} R_u^{\mathrm{W}} + \frac{h}{N}R_{\max}(h) = \frac{1}{N-h}\sum_{u=1}^{U-h} R_u^{\mathrm{W}}$$

and we need to find h such that

$$R_i^{\mathrm{W}} \le R_{\max}(h) \quad \forall\, i < U-h, \tag{20}$$
$$R_i^{\mathrm{W}} \ge R_{\max}(h) \quad \forall\, i \ge U-h. \tag{21}$$

It is easy to show that there will always exist exactly one value of $0 \le h \le N-1$ that satisfies both inequalities.

After determining h, we can find the power required by each user, and hence the sum power of the BS. Iterating over the initial power level will give the appropriate power level that matches the total power P (noting that the total power is monotonically increasing with the level of the water ρ).

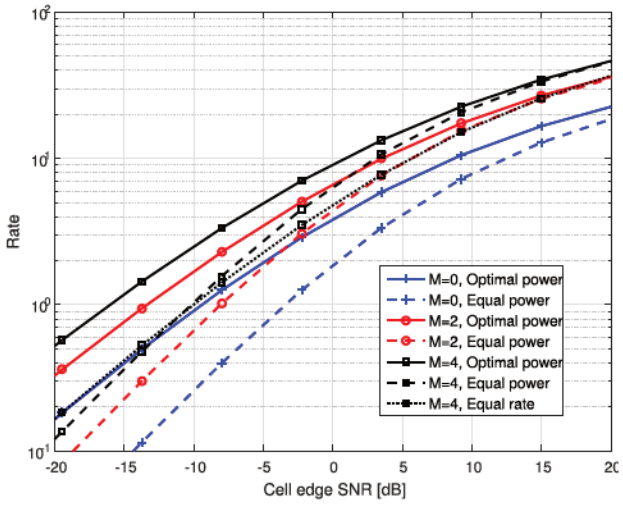

Fig. 2. Total network throughput as a function of the received power at cell edge for various cache sizes.

V. Numerical Results

In this section we present numerical results to better illustrate the gain obtained by cache-aided communication scheme and the proposed optimization framework. The performance presented here are based on Monte Carlo simulation with 1000 network realizations per point. Each Network realization consists of a random positioning of the $U = M+N$ users independently and uniformly over a circular area of radius $r_{\max}$. The channel gain were evaluate using the distances to the BS, and a random generation of $B = 100$ Rayleigh fading variables per user. In all simulations we considered single antenna users ($N_{\mathrm{R}} = 4$) and a BS with $N_{\mathrm{T}} = N = 4$ antennas.

Fig. 2 depicts the total network throughput as a function of SNR at the cell edge.[4] In this study, the throughput is defined by the data that is delivered to the users during the delivery phase (not including the data that was previously placed in their cache). The figure depicts the performance for the cases that the overall cache memory at all users contains $M = 2$ or $M = 4$ copies of the entire database. For reference, the figure also depicts the performance with no cache ($M = 0$). Following our assumption in Section IV, the number of users changes according to the allocated cache size so that $U = M+N$. The figure depicts the performance with and without caching for three types of resource allocation. 'Optimal power' depicts the performance with the optimal power allocation and optimal cache allocation as described in Section IV. 'Equal power' uses the same power for all frequency bins of all users, but optimal cache allocation. 'Equal rate' uses the power allocation that provide equal rate all users (with optimal water-filling intra-user power allocation for the different frequency bins of all users).

[4]Recall that each user experiences a different SNR. Thus, the SNR at the cell edge is a convenient reference point, even though no user is actually located at the cell edge.

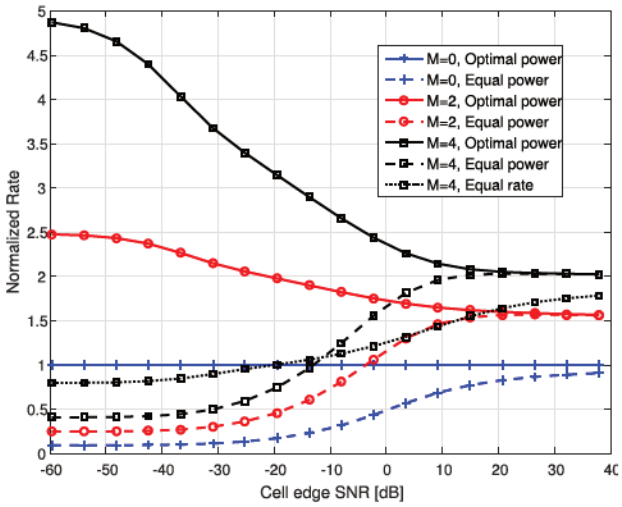

Fig. 3. Normalized network throughput as a function of the SNR at cell edge for various cache sizes.

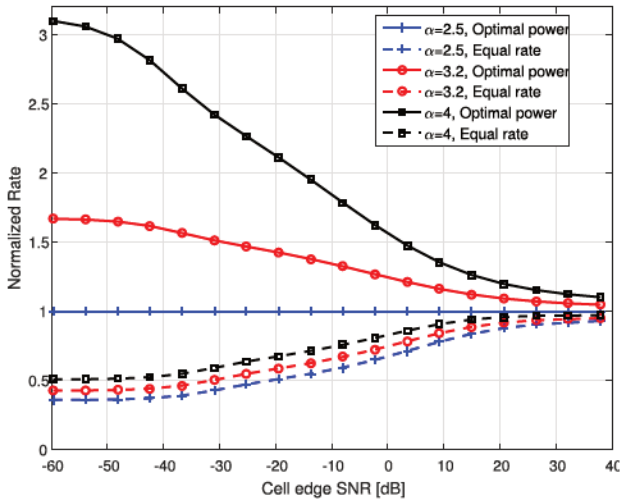

Fig. 4. Normalized network throughput as a function of the SNR at cell edge for various path loss exponents.

Note that for the 'Equal rate' the optimal cache allocation is equal for all users. This scheme characterizes the performance of previously published schemes in which the transmission rate is taken as the minimal achievable rate among the active users (e.g., [5], [30], [36]). Obviously, for such schemes were the performance is bounded by the minimal rate, the optimal power allocation leads to 'Equal rate'.

The figure shows that caching and optimal power allocation improve the performance. Yet, the fine details are hard to observe due to the large range of the vertical axis. Fig. 3 presents the same rates, but in a normalized manner that allows a better inspection. In this figure, each of the total rates was divided by the total rate in the case of optimal power allocation with no cache available at the users.

At high SNR regime, the difference between the channel gain of the different users becomes negligible, and all users approach the same rate. Hence, rate balancing is not critical and we only see the effect of throughput increase. As each scheme allows for serving $M+N$ users simultaneously, we expect a gain of $(M+N)/N$, which is 1.5 and 2 for the case of $M=2$ and $M=4$, respectively. We see that these values are indeed achieved with or without optimal power allocation. This shows that optimal power allocation has a minimal effect on the overall throughput in this regime.

On the other hand, at low SNR regime, the difference between user rates is significant, and we also see the effect of the rate balancing. It is transparent that the inherent rate balancing effect of the optimized caching scheme leads to a significant increase in the network throughput. Thus, the ability to maximize the throughput while keeping strict fairness gives gains which are close to twice the throughput gains.

Note that for $M=0$, the maximum rate constraint in (18) requires that all rates to be equal. This is reasonable as in the absence of cache, we have no balancing mechanism. Thus, in this case the powers must be set such that all users achieve exactly the same rate. Hence, the performance of the 'Optimal power' and the 'Equal rate' schemes for $m=0$ are identical. In comparison, the 'Equal power' scheme achieves lower rates due to the intra-user power allocation, i.e., the less efficient use of the frequency bins with good fading. Also note that at very low SNR, the 'Equal rate' scheme suffers a small decrease in performance with larger cache sizes. This is due to the strict fairness constraint, that requires the system to bring exactly the same rate to a larger number of users simultaneously, while the increase in DoF is meaningless at such low SNRs.

Fig. 4 shows the normalized rate for various values of the path loss exponent, α. Again, at high SNR, the rates are almost identical, and all schemes have similar performance. However, for low SNR we see significant difference between the curves. We note that the amplitude variations between the users are more considerable for larger values of the path loss exponent. Thus, a more significant gain of the balancing mechanism can be observed for larger values of α. In particular, we see the largest gain for $\alpha=4$, and the second largest gain for $\alpha=3.2$.

VI. Conclusion

In this paper we studied the cache-aided communication problem for cellular networks. An important feature of the considered model is availability of multiple antennas at the base station and the users. The links between the BS and users are assumed to be asymmetric, and are modeled by wideband fading channels. While it is known that cache and spacial diversity can be traded to achieve DoF, our analysis is not limited to DoF, and we have studied the time of delivery in finite signal-to-noise ratio regime.

We formulated the cache allocation, cache placement, and delivery scheme as a joint linear program. Even though the number of variables in the LP is large (exponential in problem parameter), the solution is very sparse (the number of non-zero variables is quadratic in problem parameters), which makes it feasible for practical implementation. The suggested scheme is better than previously known schemes as each user can be served at the rate of its own channel, rather than being compromised by the poorest user in the communication group.

We also considered a special case of the parameters for which a closed form solution can be obtained. This closed form solution was used to derive the optimal power allocation algorithm. It is shown that the joint optimization of cache usage and power allocation yields a gain for MIMO coded-caching, which goes beyond the throughput increase. In particular, it is shown that in this case, the cache is used to balance the users such that fairness and throughput are no longer contradicting. More specifically, in this case, strict fairness is achieved jointly with maximizing the network throughput.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the guest editors for their insightful comments.

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021," White Paper, 2017.
[2] (2014). *Global internet Phenomena Report: 1H 2014*. [Online]. Available: https://www.sandvine.com/downloads/general/global-internetphenomena/2014/1h-2014-global-internet-phenomena-report.pdf
[3] G. Huston, "Web caching," *Internet Protocol J.*, vol. 2, no. 3, pp. 2–20, 1999.
[4] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
[5] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj. (2017). "Multi-antenna coded caching." [Online]. Available: https://arxiv.org/abs/1701.02979
[6] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
[7] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.
[8] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4388–4413, Jul. 2017.
[9] A. Sengupta, R. Tandon, and T. C. Clancy, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1691–1695.
[10] C.-Y. Wang, S. H. Lim, and M. Gastpar, "A new converse bound for coded caching," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Jan./Feb. 2016, pp. 1–6.
[11] N. Ajaykrishnan, N. S. Prem, V. M. Prabhakaran, and R. Vaze, "Critical database size for effective caching," in *Proc. 21st Nat. Conf. Commun. (NCC)*, Feb./Mar. 2015, pp. 1–6.
[12] C. Tian. (2015). "A note on the fundamental limits of coded caching." [Online]. Available: https://arxiv.org/abs/1503.00010
[13] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2016, pp. 161–165.
[14] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr. (2016). "The exact rate-memory tradeoff for caching with uncoded prefetching." [Online]. Available: https://arxiv.org/abs/1609.07817
[15] Z. Chen, P. Fan, and K. B. Letaief. (2014). "Fundamental limits of caching: Improved bounds for small buffer users." [Online]. Available: https://arxiv.org/abs/1407.1935
[16] C. Tian and J. Chen, "Caching and delivery via interference elimination," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 830–834.
[17] S. Sahraei and M. Gastpar, "K users caching two files: An improved achievable rate," in *Proc. Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2016, pp. 620–624.
[18] M. M. Amiri and D. Gündüz, "Fundamental limits of coded caching: Improved delivery rate-cache capacity tradeoff," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 806–815, Feb. 2017.
[19] C.-Y. Wang, S. S. Bidokhti, and M. Wigger, "Improved converses and gap-results for coded caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2428–2432.
[20] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Characterizing the rate-memory tradeoff in cache networks within a factor of 2," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 386–390.
[21] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2014.
[22] S. S. Bidokhti, M. Wigger, and A. Yener, "Gaussian broadcast channels with receiver cache assignment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[23] 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 *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 1004–1008.
[24] S. Yang, K.-H. Ngo, and M. Kobayashi, "Content delivery with coded caching and massive MIMO in 5G," in *Proc. 9th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, Sep. 2016, pp. 370–374.
[25] S. S. Bidokhti, M. Wigger, and R. Timo. (2016). "Noisy broadcast networks with receiver caching." [Online]. Available: https://arxiv.org/abs/1605.02317
[26] S. S. Bidokhti, M. Wigger, and R. Timo, "Erasure broadcast networks with receiver caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1819–1823.
[27] A. Ghorbel, M. Kobayashi, and S. Yang, "Content delivery in erasure broadcast channels with cache and feedback," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6407–6422, Nov. 2016.
[28] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.
[29] K.-H. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 548–562, Jan. 2018.
[30] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Optimization of heterogeneous caching systems with rate limited links," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[31] S. S. Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1222–1226.
[32] S. S. Bidokhti, M. Wigger, A. Yener, and A. El Gamal. (2018). "State-adaptive coded caching for symmetric broadcast channels." [Online]. Available: https://arxiv.org/abs/1802.00319
[33] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "On the optimality of separation between caching and delivery in general cache networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1232–1236.
[34] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.
[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[36] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang. (2015). "The performance analysis of coded cache in wireless fading channel." [Online]. Available: https://arxiv.org/abs/1504.01452

Itsik Bergel received the B.Sc. degrees in electrical engineering and in physics from the Ben-Gurion University of the Negev, Beer-Sheva, Israel, in 1993 and 1994, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Tel Aviv, Tel-Aviv, Israel, in 2000 and 2005, respectively. From 2001 to 2003, he was a Senior Researcher with the Intel Communications Research Laboratory. In 2005, he was a Post-Doctoral Researcher with the Dipartimento di Elettronica, Politecnico di Torino, Italy. He is currently a Faculty Member with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel. His main research interests include multichannel interference mitigation in wire line and wireless communications, cooperative transmission in cellular networks, and cross layer optimization of random ad-hoc networks.

Soheil Mohajer received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Iran, in 2004, and the M.Sc. and Ph.D. degrees in communication systems from the École Polytechnique Fédérale de Lausanne, Switzerland, in 2005 and 2010, respectively. He was a Post-Doctoral Researcher with Princeton University from 2010 to 2011, and the University of California at Berkeley, from 2011 to 2013. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. He is broadly interested in information theory, wireless networks, distributed storage systems, and statistical machine learning. He received the NSF CAREER Award in 2018.