Simple Caching Schemes for Non-homogeneous MISO Cache-Aided Communication via Convexity

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

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

***Abstract*—We present a novel scheme for cache-aided communication over multiple-input and single output (MISO) cellular networks. The presented scheme achieves the same number of degrees of freedom as known coded caching schemes, but, at much lower complexity. The scheme is derived for communication systems with heterogeneous rates and finite signal-to-noise ratio, in which links are modeled by wideband fading channels. The base station is serving multiple users simultaneously, by sending a combination of several packets, each intended for one user. The interference is either suppressed using the cache content or nulled by zero-forcing at the unintended users. We focus on *efficient* coding schemes, which allow for a maximum number of users to be served throughout the course of communication. An achievable rate region is characterized by determining the extreme rate vectors satisfying an efficient transmission. The analysis results in a simple scheduling scheme and in a closed-form performance analysis.**

***Index Terms*—Cache-aided communication, MISO, Finite SNR regime, Zero-forcing, Simple scheduling.**

I. Introduction

Wireless communication technologies and data delivery networks have been improved significantly over the recent years. However, with the dramatic growth in the popularity of large data and high speed applications, the rates supported by these networks are not likely to keep up with the demand. cache aided communication is a promising strategy to overcome this challenge by better exploiting the network resources during their peak-traffic time.

Cache aided communication (sometimes referred to as coded caching) operates in two phases: placement phase and delivery phase [1]. During the placement phase, we can prefetch and store at each users' memory some packets from the files in the database, without the actual request of the user being known. Once the requests are revealed, the server generates a set of joint messages and transmits them to all the receivers during the delivery phase. These messages should be formed such that each user will be able to decode its desired file from its cache content and the received signal.

While the gain of conventional caching schemes is limited to the fraction of the database stored at each individual user (which is negligible in practice), the 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. In [1], authors considered the single shared-link network and proposed a coding strategy with a significant caching gain, which was later shown to be optimal [2].

Spatial diversity is another resource in wireless networks which offers a significant improvement in the network throughput by deploying multiple antennas at the transmitter and/or receivers. Interestingly, the spatial diversity gain and caching gains can be simultaneously achieved: It is shown in [3] that in a broadcast system with N transmit antennas at the server and an aggregate cache size that can distributedly store M complete copies of the database, a total of $N + M$ degrees of freedom (DoF) can be achieved. In spite of the current trend of massive MIMO, this gain is substantial, especially since a small cache at each mobile comes at a very low cost, and these memories easily scale up with the number of users.

The main problem with the approach in [3] is its complexity, also known as the sub-packetization problem. Denoting the number of user by U, this scheme requires $\binom{U}{N+M}$ transmissions. Furthermore, each transmission requires each user to decode $\binom{N+M+1}{M}$ messages. As cache aided communication is attractive mostly for networks with large number of users, the resulting complexity limited the practicality of this scheme.

This scheme was further extended to consider the effect of fading [4], [5] and the networks with heterogeneous rates [6]. Power allocation was also considered, but typically as an isolated optimization problem [7]–[10]. Yet, all these variants shared the same complexity order, and hence could not lead to practical implementation.

A simpler transmission scheme was considered in [11], where the BS transmits multiple packets simultaneously to $N+M$ users using the N antennas. Each message is decoded at a single user, where the BS uses zero-forcing (ZF) precoding to avoid the interference between some of the messages, while users use their cache contents to remove the remaining interference. This approach achieves the same number of DoFs as [3] and can also support heterogeneous rates. More importantly, this scheme has much simpler transmissions, where each user receives only a single message at each transmission.

The downside of [11] is that the optimal transmission scheme is not given in closed form, but only as the solution of an optimization problem. Furthermore, the number of variables

in this optimization problem grows exponentially with the number of users, so that the complexity issue is not truly resolved,

In this work, we present a straight forward transmission scheme that can harness the advantages of cache aided communication at a very low complexity. This scheme enjoys the simplicity of the transmission scheme of [11], but avoids the optimization problem. We also characterize in closed form the rate set that is supported by this scheme. This characterization allows an analytic performance analysis, and also an optimization of the power allocated to each user.

II. System model

We consider a single cell network with one base station (BS), equipped with N transmit antennas, and U users each with one receive antenna. The BS has access to a dictionary of D files, where $D \geq U$. Each file has a normalized size of 1. Each user has a memory of size $q = MD/U$, which will be filled in a *centralized* fashion by (uncoded) packets of the files during the placement phase. Thus, we can prefetch and store a total of M copies of the entire dictionary, distributedly across the users.

We focus on 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

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

where $\mathbf{x}_m \in \mathbb{C}^{N\times 1}$ is the transmit vector, $w_{i,m} \sim CN\left(0, \frac{BN_0}{F}\right)$ is the additive noise sample, N_0 is the power spectral density of the 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

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

where α is the path-loss exponent, $\mathbf{g}_{i,m} \sim CN(0, \mathbf{I})$ represents the random fading, and r_i is the distance between the i-th mobile and the BS.

The BS sends a combination of different messages at each time instance, to simultaneously serve multiple users. Some of these messages can be ignored at the mobile 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.

More precisely, the BS serves $M+N$ users simultaneously at any given time (over all frequency bins). Optimal performance would require optimization of the power allocated to each frequency bin of each user, subject to a BS average power constraint of P Watts per frequency bin. Such optimization is left for the full version of this work.

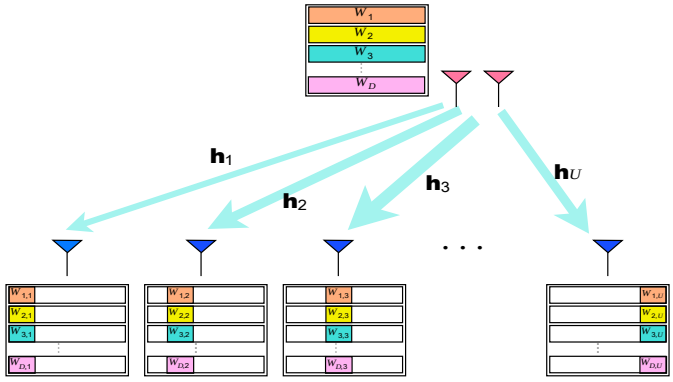

Fig. 1: The system model with $N = 2$ transmit antennas.

To avoid interference, the precoding vector for each message at each frequency bin, $\mathbf{f}_{i,m} \in \mathbb{C}^{N\times 1}$, 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 the transmission of this message. All other users are either not active during this symbol, or have this undesired message in their cache, and can subtract it from their received signal.

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 user rate, $\tilde{R}_i$, converges to its expectation:

$$\tilde{R}_i \to R_i = B \cdot \mathbb{E}\left[\log_2\left(1 + \frac{P \cdot r_i^{-\alpha}}{BN_0(M+N)}\eta_{i,m}\right)\right], \qquad (3)$$

where N_0 is the noise spectral density at the users, $\eta_{i,m} = |\mathbf{g}_{i,m}\, \mathbf{f}_{i,m}|^2$, and B/F is the symbol rate. Setting $\|\mathbf{f}_{i,m}\|^2 = 1$, and recalling the complex normal distribution of $\mathbf{g}_{i,m}$, we note that $\eta_{i,m}$ admits an exponential distribution with a mean of 1.

For simplicity, we focus in this work on the simpler case of $M = 1$, and leave the treatment of larger cache sizes to future study. Thus, we have one copy of the database cached across the network, and the centralized placement consists of dividing each file into U equal size sections and storing the i-th section of all files at the cache of User i.

III. Transmission scheme and the achievable rate region

As in [11] we focus on *efficient* transmission schemes as formally defined below:

Definition 1: A delivery scheme is called *efficient* if the base-station serves $M+N$ users at any given time of the transmission. A rate vector $\mathbf{R} = [R_1, \ldots, R_U]$ is said to be *efficiently achievable* if there exists an efficient transmission scheme for this vector.

It was shown in [11] that any efficient scheme can serve one file request for each user within a total transmission time of

$$T = \frac{U-M}{(N+M)U} \sum_k \frac{1}{R_k}. \qquad (4)$$

In the rest this work we focus on characterizing the efficiently achievable rate vectors, and in particular, we obtain

some sufficient conditions for a rate vector $\mathbf{R}$ to be efficiently achievable. Moreover, we devise an explicit and low-complexity transmission scheme for the efficiently achievable rate vectors identified in this paper. These questions where answered in [11] only through the solution of a large optimization problem. In this paper we provide an insightful and intuitive solution for these questions. Specially, we show that the desired transmission scheme for a specific rate vector can be obtained by a linear combination of simpler transmission schemes.

We adopt the notation of [11] to describe the transmission scheme. In particular, for a combination of $N+1$ active users denoted by c and the j-th transmission slot, we use the binary $U \times U$ matrix $\mathbf{E}_j^c$ to define the transmission scheme, where $(\mathbf{E}_j^c)_{i,\ell} = 1$ if the i-th user receives (part of) the ℓ-th section of its requested file, and otherwise $(\mathbf{E}_j^c)_{i,\ell} = 0$. Thus, we have $1 \le i \le U$, $1 \le \ell \le U$ and $0 \le c \le \binom{U}{N+1}$. As we focus on efficient transmission schemes, all matrices satisfy: $\sum_\ell \sum_i (\mathbf{E}_j^c)_{i,\ell} = M+N$.

Denoting by T_j^c the time required to transmit to a user combination c in mode j, the total transmission time is:

$$T = \sum_{j,c} T_j^c. \tag{5}$$

Furthermore, for a rate vector $\mathbf{R} = [R_1, \ldots, R_U]$ we define the equivalent *delay vector* as $\mathbf{A} = [1/R_1, \ldots, 1/R_U]$. Now, recall that each user has a cache to store $1/U$ fraction of the dataset. Hence, a delay vector $\mathbf{A}$ is efficiently achievable if there exists a duration vector $\mathbf{T} = \{T_j^c\}$ such that

$$\sum_{j,c} T_j^c \cdot (\mathbf{E}_j^c)_{i,\ell} = \frac{A_i}{U}, \quad \forall i \neq \ell. \tag{6}$$

The following lemma uses the convexity of the problem and its scalability with a positive constant to show that any linear combination of two efficiently achievable delay vectors with positive coefficients is efficiently achievable.

Lemma 1: Let $\mathbf{A}_1 = [A_{1,1}, \ldots, A_{1,U}]$ and $\mathbf{A}_2 = [A_{2,1}, \ldots, A_{2,U}]$ be two efficiently achievable delay vectors corresponding to the optimal schedules $\mathbf{T}_1 = \{T_{1,j}^c\}$ and $\mathbf{T}_2 = \{T_{2,j}^c\}$, respectively. Then for any $c_1, c_2 \ge 0$ with $c_1 + c_2 > 0$ the vector $\mathbf{A} = c_1\mathbf{A}_1 + c_2\mathbf{A}_2$ is efficiently achievable, with a corresponding schedule $\mathbf{T} = c_1\mathbf{T}_1 + c_2\mathbf{T}_2$.

Proof: First note that since $c_1, c_2 \ge 0$, we obviously have $T_j^c = c_1 T_{1,j}^c + c_2 T_{2,j}^c \ge 0$. Thus, proving achievability only requires proving that

$$\sum_{j,c} T_j^c \cdot (\mathbf{E}_j^c)_{i,\ell} = \frac{A_i}{U} \tag{7}$$

for all distinct values of ℓ and i. Substituting, we get

$$\begin{aligned}\sum_{j,c} T_j^c \cdot (\mathbf{E}_j^c)_{i,\ell} &= c_1 \sum_{j,c} T_{1,j}^c \cdot (\mathbf{E}_j^c)_{i,\ell} + c_2 \sum_{j,c} T_{2,j}^c \cdot (\mathbf{E}_j^c)_{i,\ell} \\ &= \frac{c_1 A_{1,i}}{U} + \frac{c_2 A_{2,i}}{U} = \frac{A_i}{U},\end{aligned} \tag{8}$$

which proves the lemma. ∎

Thus, in order to characterize the desired transmission scheme, it suffices to determine a set of *edge delay vectors*, since the achievability of all interior delay vectors is implied by Lemma 1. Each edge vector is obtained by having one weak user and all other users with identical rates. We denote by $\mathbf{1}$ the vector of all ones, and by $\mathbf{i}_u$ the vector in which the u-th element is 1 and all other elements are 0. The next lemma determines the set of edge vectors.

Lemma 2: The delay vector $\mathbf{A}_u = \mathbf{1} + \frac{U-N-1}{N} \cdot \mathbf{i}_u$ is efficiently achievable.

Proof: This proof requires finding a scheduling duration vector $\mathbf{T}_u = \{T_{u,j}^c\}$ that satisfies (7). Due to space constraints, this proof is relegated to the journal version of this work. Here we'll just mention that the transmission scheme is comprised of two sets of transmissions. In one set $N+1$ users are served, where each user uses its cache to cancel interference from one message intended to another user. In the second set, two users simultaneously use their cache to cancel interference from the message sent to the weak user. ∎

The following theorem states whether a rate vector is efficiently achievable using our scheme, and gives the desired transmission scheme:

Theorem 1: Any rate vector $\mathbf{R} = [R_1, \ldots, R_U]$ that satisfies

$$\max_k R_k \le \frac{1}{N} \cdot \frac{(N+1)(U-1)}{\sum_k \frac{1}{R_k}} \tag{9}$$

is efficiently achievable, and files requested by all users can be delivered using the allocation vector:

$$\mathbf{T} = \sum_u c_u \mathbf{T}_u \tag{10}$$

where

$$c_k = \frac{N}{U-N-1} A_k - \frac{N}{(N+1)(U-1)} \sum_\ell A_\ell \cdot \tag{11}$$

Proof of Theorem 1: By a straight-forward generalization of Lemma 1 to multiple delay vectors and applying it to the edge vectors defined in Lemma 2 we conclude that any delay vector of the form $\mathbf{A} = \sum_u c_u \mathbf{A}_u$ with $c_u > 0$ is achievable. Let $\mathbf{A} = [A_1, \ldots, A_U]$ and $\mathbf{A}_u = [A_{u,1}, \ldots, A_{u,U}]$ for $u = 1, \ldots, U$ To characterize the achievable region, we first note that the sum of all the entries of the delay vector must satisfy

$$\sum_\ell A_\ell = \sum_u c_u \sum_\ell A_{u,\ell} = U + \frac{U-N-1}{N} \sum_u c_u. \tag{12}$$

Now, considering the delay of the k-th user, we have

$$A_k = \sum_u c_u A_{u,k} = \frac{U-N-1}{N} c_k + \sum_u c_u, \tag{13}$$

and combining with (12) gives (11).

This is a feasible solution as long as $c_k \ge 0$ for all k. Thus, (13) results in the sufficient condition

$$\min_k A_k \ge \frac{N}{(N+1)(U-1)} \sum_k A_k \tag{14}$$

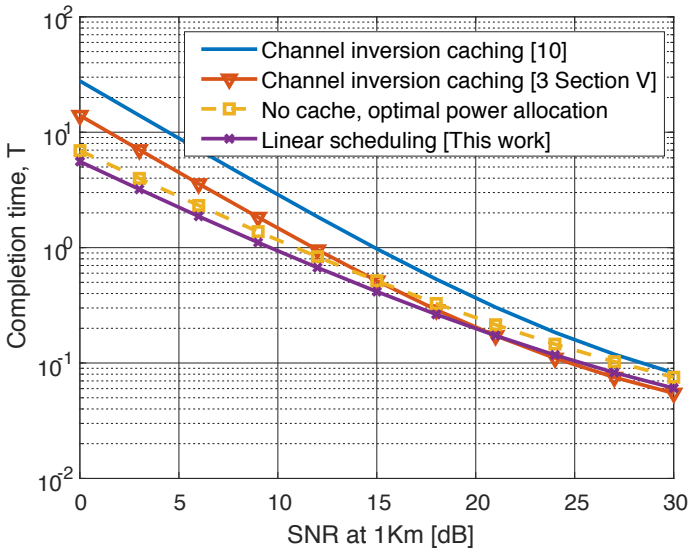

Fig. 2: Completion time vs. SNR for various caching schemes, with and without optimal power allocation.

which translates to the condition in (9), by replacing $A_u = 1/R_u$. This completes the proof of the theorem. ■

IV. Scheme Complexity

The main complexity issue in cache aided communications is the number of sub-packets that need to be transmitted to serve all users. For example, the scheme in [3] for $M = 1$ requires transmission of $U\binom{U-2}{N-1}$ packets (i.e., approximately $U^N/(N-1)!$). The optimization scheme of [11] guarantees a solution with at most $(N+1)U^2$ packets. But this solution still requires solving of an optimization problem, whose size grows very fast with the number of users. The proposed scheme can take advantage of the scheme of [11] to find one of the edge vectors, and keep it in a lookup table. Then the generation of an appropriate transmission scheme for a given set of user rates is straight forward using Theorem 1. The number of transmitted packets is upper bounded by $(N+1)U^3$ (a linear combination of at most U edge delay vectors, where each edge delay vector can require the transmission of at most $(N+1)U^2$ packets [11]). Thus, we have a significant reduction in complexity compared to [3] in systems with more than $N = 3$ antennas. Furthermore, the complexity of the proposed scheme depends only linearly as the number of antennas, and also can support non-homogeneous rates.

V. Numerical Results

In this section we present numerical results that illustrate the achievable rates using our proposed scheme. We emphasize again that the main advantage of this scheme is its complexity, which depends only linearly on N. As an example, we consider in this section a system with $U = 30$ users and $N = 5$ antennas. The scheme in [3] would require the transmission of $614,250$ packets, while our scheme will require about one quarter of that. This difference will become even more emphasized for a larger number of users or more antennas. (It will also become more emphasized for larger cache sizes, $M > 1$, but this is left for future study.)

Our scheme inherits from [11] a lower power efficiency compared to [3]. This is due to the simple addition of messages instead of xoring. But, our scheme also brings a significant advantage. Not only that it can support different rates for different users, we also have a closed form characterization of its achievable rate region. This allows a simpler derivation of optimal power allocation schemes. Due to lack of space, we relegate the derivation of the power allocation scheme to the journal version of this work, and only report here the final results.

To demonstrate the importance of combining caching and power allocation, Fig. 2 depicts the average time needed to serve all 30 users in a cell with 5 transmit antennas. The users are located randomly, where the distance of each user to BS has an exponential distribution with a mean of 1Km. The revived power at each mobile is determined by a path loss exponent of 4, and the x-axis gives the average SNR at a distance of 1Km. The file length is 10^7 bits and the transmission is performed over a wideband channel with a bandwidth of 10MHz, which is divided into 1000 frequency bins of equal bandwidths. The transmission experiences fast fading with a coherence bandwidth much smaller than the transmission bandwidth.

All cashing schemes in the figure have the same number of DoF, and hence will require the same asymptotic transmission delay at high enough SNR. At low SNR, the channel gains and power allocation play a significant role, and the effect of good power allocation is clearly visible.

As optimal power allocation is not known for the schemes of [3] and [10], we use a channel inversion power allocation and assign equal power to all frequency bins. Thus, these scheme operate with equal rates for all users. This is shown to be a major drawback at low SNR, where our novel scheme shows a significant advantage.

VI. Conclusions

We presented a novel scheme for cache-aided communication over MISO cellular networks. The scheme has lower complexity than previously published schemes, and supports inhomogeneous user rates. Furthermore, we presented a closed form expression for the achievable rate region of the presented scheme, which enables the derivation of optimal power allocation. Numerical results show that the optimal power allocation becomes a significant advantage at low SNRs, when the difference between user rates is most significant. Thus, apart from its lower complexity, this novel scheme can also give a performance advantage at low SNR.

The journal version of this work provides the missing proof for the achievability of the edge delay vectors. It also gives the derivation of the optimal power allocation scheme (which was used for the numerical results) and the extension of the results for larger cache sizes ($M > 1$). Further research is still needed to improve the energy efficiency of this scheme compared to the energy efficient xoring operation in [3].

REFERENCES

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. 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," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1281–1296, 2018.

[3] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Multi-antenna coded caching," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2113–2117.

[4] 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.

[5] 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, 2016.

[6] 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.

[7] S. S. Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," in *IEEE Int. Symp. Inf. Theory*, 2017, pp. 1222–1226.

[8] 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.

[9] M. M. Amiri and D. Gündüz, "Decentralized caching and coded delivery over gaussian broadcast channels," in *IEEE Int. Symp. Inf. Theory*, 2017, pp. 2785–2789.

[10] S. Mohajer and I. Bergel, "Optimal power allocation in miso cache-aided communication," in *2018 IEEE 19th Int. Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2018.

[11] I. Bergel and S. Mohajer, "Cache aided communications with multiple antennas at finite SNR," *IEEE J. Sel. Areas Commun.*, 2018.