Subpacketization-Beamformer Interaction in Multi-Antenna Coded Caching

MohammadJavad Salehi*, Antti Tölli*, Seyed Pooya Shariatpanahi†
*Center for Wireless Communications, University of Oulu, Oulu, Finland.
†School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran.
{fist_name.last_name}@oulu.fi; p.shariatpanahi@ut.ac.ir

***Abstract*—We study the joint effect of beamformer structure and subpacketization level on the achievable rate of cache-enabled multi-antenna communications. We use appropriate low-SNR approximations, to show that using simple zero-forcing (ZF) beamformers, increasing subpacketization degrades the achievable rate; in contrast to what has been shown in the literature for more complex, optimized beamformers. We also numerically analyze the probability distribution of symmetric rate terms, in order to confirm the validity of mathematical outputs. The results suggest that for improving the content delivery rate at low-SNR, subpacketization level and beamformer complexity should be jointly increased.**

***Index Terms*—Coded Caching, Multi-Antenna Communications, Subpacketization, Beamformer Design**

I. Introduction

Network data volume has continuously grown during the past years. The global IP (Internet Protocol) data volume is expected to exceed 4.8 Zettabytes (10^{21} bytes) by 2022, from which 71 percent will pass through wireless networks [1]. On the other hand, introduction of new application types for 5G and beyond (e.g. autonomous vehicles, immersive viewing and massive machine-type communications) has necessitated extreme advancements for all networking KPIs (Key Performance Indicators) such as data rate, delay and reliability [2]. This has imposed serious challenges in all data network layers, and solving them is one of the main recent research trends.

Coded Caching (CC), recently proposed in [3], is considered as a solution to higher data rate requirements, specially for the prominent use-case of video-based applications. The idea is to transmit carefully designed codewords over shared data links; enabling an additional global caching gain, proportional to the total cache size in the network, to be achieved in addition to the local caching gain at each node. Interestingly, it is shown that CC gain is additive with the multi-antenna gain [4], making it even more desirable for next-generation wireless networks.

Coded caching has been studied extensively in the literature. A major part of the studies is dedicated to the information-theoretic analysis; i.e. finding the maximum possible CC gain under various assumptions [5], [6], as well as designing less complex solutions for achieving a near-optimal performance [7], [8]. Specifically, reducing subpacketization, defined as the number of smaller parts each file should be split into, is considered for various CC setups; and is shown to be nicely achievable in cache-enabled multi-antenna networks [7], [8].

The aforementioned studies consider error-free communication links for all users. From another perspective, CC performance at low-SNR is also addressed in the literature. In [9] it is shown that the achievable rate of CC at low-SNR can be improved considerably, by using optimized beamformers instead of zero-forcing (ZF). In [10] a flexible-subpacketization CC design is introduced and it is shown that the achievable rate is improved, as subpacketization is increased. The results are only valid for the case of optimized beamformers though.

In this paper, we extend the results of [10], by studying the joint effect of subpacketization level and beamformer structure on low-SNR performance. We show that the positive effect of increased subpacketization on the achievable rate might vanish (or even be reversed), if simpler ZF beamformers are used instead of optimized ones. This provides a deeper understanding of how the structure complexity affects the performance of multi-antenna CC setups; enabling CC schemes to be better tailored for real-world implementations.

Throughout the text, we use $[K]$ to denote $\{1, 2, ..., K\}$ and $[i : j]$ to represent $\{i, i+1, ..., j\}$. Boldface upper- and lower-case letters denote matrices and vectors, respectively. $\mathbf{V}[i, j]$ refers to the element at the i-th row and j-th column of matrix $\mathbf{V}$. Sets are denoted by calligraphic letters. For two sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A}\backslash\mathcal{B}$ is the set of elements in $\mathcal{A}$ which are not in $\mathcal{B}$; and $|\mathcal{A}|$ represents the number of elements in $\mathcal{A}$.

II. System Model and Literature Review

A. Problem Setup

We consider a multiple input, single output (MISO) setup. A server, equipped with L transmitting antennas, communicates with K single-antenna users over a shared wireless link. The server has access to the file library $\mathcal{F}$, which has N files each with size f bits. Every user $k \in [K]$ has a cache memory of size Mf bits, where $M \leq N$. For simplicity, we use a normalized data unit and drop f in subsequent notations.

The system operation consists of two phases; placement and delivery. At the placement phase, cache memories of the users are filled with data from the files in $\mathcal{F}$. This is done without any knowledge of file request probabilities in the future; and hence an efficient strategy is to store equal-sized data portions of all files (with size M/N) in the cache memory of each user. We use $\mathcal{Z}(k)$ to denote the cache contents of user k.

This work was supported by the Academy of Finland under grants no. 319059 (Coded Collaborative Caching for Wireless Energy Efficiency) and 318927 (6Genesis Flagship).

At the beginning of the delivery phase, every user k reveals its requested file $W(k) \in \mathcal{F}$. Consider $\mathcal{D} = \{W(k) \mid k \in [K]\}$ to be the demand set. Based on $\mathcal{D}$ and $\mathcal{Z}(k)$, the server builds and transmits (e.g. in a TDMA fashion) transmission vectors $\mathbf{x}(i) \in \mathbb{C}^L, i \in [I]$; where I is a parameter determined by the delivery algorithm. After $\mathbf{x}(i)$ is transmitted, user k receives $y_k(i) = \mathbf{h}_k^T \mathbf{x}(i) + z_k(i)$, where $\mathbf{h}_k \in \mathbb{C}^L$ is the channel vector for user k and $z_k(i) \sim \mathbb{C}\mathcal{N}(0, N_0)$ represents the observed noise at user k during transmission interval i.

As M/N of every file is available at $\mathcal{Z}(k)$, user k needs to get $(1 - M/N)$ of $W(k)$ from the server. Let us define the CC gain as $t = KM/N$ and assume t is an integer. Defining delivery time T as the time required for all users to decode their requested files, the effective communication rate is $R = K(1-t/K)/T$. For a simple channel with capacity one data unit per channel use, R represents how many users simultaneously benefit from each transmission. We use the term Degree of Freedom (DoF) equivalent to R in such a case.

If $L = 1$, a trivial scheme (unicasting every missing data part) achieves DoF of one. Interestingly, CC enables DoF of $t+1$ to be achieved, for the same setup [3]. During placement phase, each file W is split into $\binom{K}{t}$ smaller parts $W_{\mathcal{T}}$, where $\mathcal{T} \subseteq [K]$, $|\mathcal{T}| = t$; and $\mathcal{Z}(k) = \{W_{\mathcal{T}} \mid W \in \mathcal{F}, \mathcal{T} \ni k\}$. For delivery, for every $\mathcal{S} \subseteq [K], |\mathcal{S}| = t + 1$, $X(\mathcal{S})$ is built as

$$X(\mathcal{S}) = \bigoplus_{k \in \mathcal{S}} W_{\mathcal{S}\backslash\{k\}}(k) \ , \tag{1}$$

where $\oplus$ denotes the bit-wise XOR operation; and each $X(\mathcal{S})$ is broadcast in a separate time interval. Following the same structure, in [11] it is shown that in a multi-server scenario with L servers, the DoF of $t + L$ is achievable. This result is then extended in [4] to an L-antenna MISO setup.

B. The Subpacketization Effect

Subpacketization P is defined as the number of smaller parts each file should be split into. The original CC scheme [3] requires $P = \binom{K}{t}$; which means P grows exponentially with K, if t also scales with K (polynomially, if t is fixed). For the multi-server scheme [11], the growth in P is even worse by a multiplicative factor of $\binom{K-t-1}{L-1}$. This makes the CC implementation infeasible, even for moderate values of K [7].

Interestingly, multi-antenna setups enable huge reductions in P to be achieved, without any loss in DoF. In [7] a structure for *elevating* single-antenna CC schemes for multi-antenna setups is introduced. The resulting CC scheme would then require $P' = g(\frac{K}{L}, \frac{t}{L})$; if for the original one $P = g(K, t)$ for some function g. However, this structure incurs DoF loss (up to a factor of 2), if either K/L or t/L is non-integer. In [8] a CC scheme with $P = K(t + L)$ is introduced, for networks with $L \geq t$. In [10] it is shown that if $K = t+L$, P can be selected freely among a set of predefined values.

C. The Beamformer Effect

Zero-forcing beamformers (ZF), used in the original multi-antenna CC scheme [4], are shown in [9] to result in poor rate in low-SNR communications. In [12] interesting methods for reducing optimized beamformer design complexity are proposed. They incur either DoF loss or increased P, however.

D. Subpacketization-Beamformer Interaction

Following [10], we assume $K = t+L$; and build $\mathcal{Z}(k)$ using a placement matrix $\mathbf{V}$, which is a $P \times K$ binary matrix with $\sum_p \mathbf{V}[p,k] = t, \forall k \in [K]$ and $\sum_k \mathbf{V}[p,k] = Pt/K, \forall p \in [P]$. Each file W is split into P smaller parts W_p; and

$$\mathcal{Z}(k) = \{W_p \mid \mathbf{V}[p,k] = 1; \ \forall p \in [P], \forall W \in \mathcal{F}\} \ . \tag{2}$$

For delivery, for each $\mathcal{S} \subseteq [K], |\mathcal{S}| = t + 1$ we define

$$\Phi(\mathcal{S}) = \{p \in [P] \mid \mathbf{V}[p,k] = 0, \ \forall k \in [K]\backslash\mathcal{S}\} \ , \tag{3}$$

and build codeword $X(\mathcal{S})$ as

$$X(\mathcal{S}) = \bigoplus_{\substack{k \in \mathcal{S} \\ p \in \Phi(\mathcal{S})}} (1 - \mathbf{V}[p,k]) W_p(k) \ . \tag{4}$$

As $K = t + L$, one can send all codewords $X(\mathcal{S})$ in a single interval. The transmission vector is built as $\mathbf{x} = \sum_{\mathcal{S}} \mathbf{w}_{\mathcal{S}} X(\mathcal{S})$; where $\mathbf{w}_{\mathcal{S}} \in \mathbb{C}^L$ is the beamforming vector associated to $X(\mathcal{S})$. Let us denote the total power constraint as P_{T}; and the number of $\mathcal{S}$ sets for which $\Phi(\mathcal{S}) \neq \varnothing$ as $n(\mathbf{V})$. We study three different beamformer structures:

- **EP:** ZF with uniform power allocation, for which

$$\mathbf{w}_{\mathcal{S}} = \sqrt{\frac{P_{\mathrm{T}}}{n(\mathbf{V})}} \times \mathbf{u}_{\mathcal{S}} \ , \tag{5}$$

  where $\mathbf{u}_{\mathcal{S}}$ is the ZF vector associated with $\mathcal{S}$, built such that $\|\mathbf{u}_{\mathcal{S}}\| = 1$ and $\mathbf{h}_k^T \mathbf{u}_{\mathcal{S}} = 0$, $\forall k \in [K]\backslash\mathcal{S}$.
- **PL:** ZF with optimal power loading, for which

$$\mathbf{w}_{\mathcal{S}} = \sqrt{\alpha_{\mathcal{S}} P_{\mathrm{T}}} \times \mathbf{u}_{\mathcal{S}} \ , \tag{6}$$

  where $\alpha_{\mathcal{S}}$ is the power coefficient of $X(\mathcal{S})$, selected such that R is maximized and $\sum_{\mathcal{S}} \alpha_{\mathcal{S}} = 1$.
- **OB:** optimized beamformer; for which

$$\mathbf{w}_{\mathcal{S}} = \sqrt{\alpha_{\mathcal{S}} P_{\mathrm{T}}} \times \mathbf{v}_{\mathcal{S}} \ , \tag{7}$$

  where $\mathbf{v}_{\mathcal{S}}$ is the beamformer vector of $X(\mathcal{S})$, designed such that R is maximized and $\|\mathbf{v}_{\mathcal{S}}\| = 1$.

According to [10], after $\mathbf{x}$ is transmitted, all unwanted terms at user k are either zero-forced (suppressed, in case of OB), or removed by $\mathcal{Z}(k)$. Consequently, user k has to decode its requested data parts from a multiple access channel (MAC) of size $m(\mathbf{V}) = P - \frac{Pt}{K} = \frac{PL}{K}$. Let us use r_k^j to denote the rate associated with the j-th term, $j \in [m(\mathbf{V})]$, at the MAC channel of user k. All terms in the MAC channel should be decoded, for user k to receive $W(k)$ successfully. This means $r_k = \min_j r_k^j$, where r_k is the perceived rate at user k. Let us define the symmetric rate as $r_s = \min r_k, k \in [K]$; and use SINR_k^j to denote the *Signal to Interference plus Noise Ratio* of r_k^j, at user k. Then the rate optimization problem is

$$\begin{aligned} \max \quad & r_s = \min_{k \in [K]} \min_{j \in [m(\mathbf{V})]} r_k^j \\ s.t. \quad & \sum_{j \in \mathcal{J}} r_k^j \leq \log(1 + \sum_{j \in \mathcal{J}} \mathrm{SINR}_k^j) \\ & \qquad \forall k \in [K]; \ \forall \mathcal{J} \subseteq [m(\mathbf{v})], \mathcal{J} \neq \varnothing \ . \end{aligned} \tag{8}$$

Note that SINR_k^j depends on the beamformer structure; and hence depending on the selected structure, it implicitly includes the power constraint ($\sum_{\mathcal{S}} \alpha_{\mathcal{S}} = 1$) and optimization variables ($\alpha_{\mathcal{S}}$ and $\mathbf{v}_{\mathcal{S}}$). To compare various schemes, we use

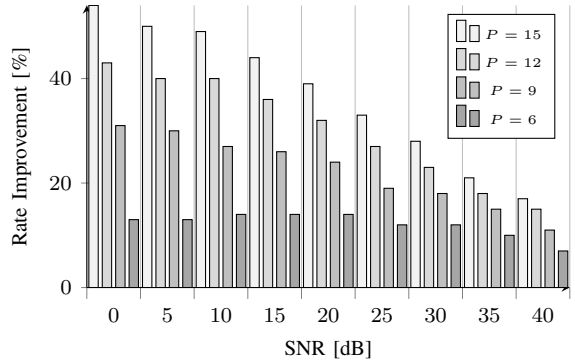

Fig. 1: Rate Improvement over $P = 3$ - OB beamformer

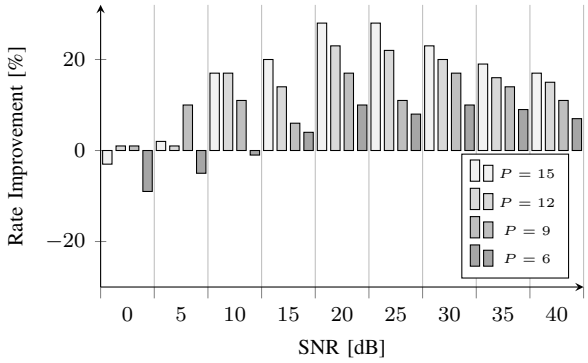

Fig. 2: Rate Improvement over $P = 3$ - PL beamformer

the total delivery time T. As the size of each data part is $1/P$ and all parts are decoded simultaneously, we have $T = \frac{1}{P}\frac{1}{r_s}$.

E. Motivation

The results of [10] indicate that using OB structure, R is improved as P is increased; and this effect is more prominent at lower SNR. The improvement in R for various P values with respect to $P = 3$ is depicted in Figure 1, for a network with $K = 6$, $t = 2$ and $L = 4$. Consequently, for the same setup, we have plotted the results for PL and EP structures, in Figures 2 and 3 respectively. Clearly, at low-SNR regime, the performance of $P > 3$ is worse than $P = 3$ for both structures. Specially for EP, at 0dB the performance is degraded as P is increased from 6 to 15. This indicates a difference in how P affects the performance, for various beamformer designs.

III. Mathematical Analysis

Using the low-SNR (e.g. SNR $\leq$ 10dB) approximation $\log(1+\text{SINR}) \simeq \text{SINR}$ (using Taylor expansion of $\log(1+x)$, as $x \to 0$; log is the natural logarithm), we can reduce (8) to

$$\begin{aligned} \max \quad & r_s = \min_{k\in[K]} \min_{j\in[m(\mathbf{V})]} r_k^j \\ s.t. \quad & r_k^j \leq \text{SINR}_k^j \qquad \forall k \in [K];\ \forall j \in [m(\mathbf{V})] \ . \end{aligned} \tag{9}$$

Note that as P becomes larger, number of constraints being removed as a result of approximation grows exponentially; and hence the approximation effect on r_s is more prominent.

We first analyze a simple network with $K = 4, t = 2, L = 2$ and then study more general setups. For the simple network, we calculate T for $P \in \{2, 4, 6\}$ and beamformer structures

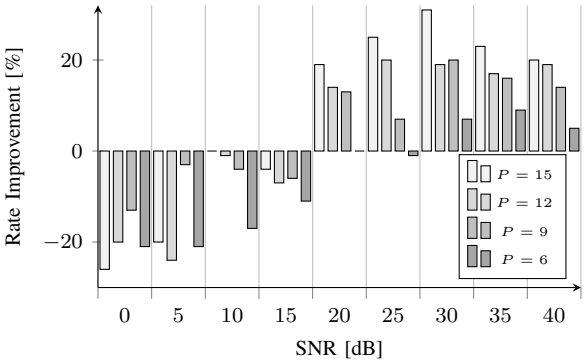

Fig. 3: Rate Improvement over $P = 3$ - EP beamformer

EP, PL, OB. Placement matrices for $P = 2, 4$ are $\mathbf{V}_1$ and $\mathbf{V}_2$ as mentioned in (10); and for $P = 6$, $\mathbf{V}_3$ is column-wise concatenation of $\mathbf{V}_1$ and $\mathbf{V}_2$.

$$\mathbf{V}_1 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \mathbf{V}_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} . \tag{10}$$

We also assume $\mathcal{D} = \{A, B, C, D\}$; and use $\mathbf{u}_k \equiv \mathbf{u}_{[4]\setminus\{k\}}$, $\mathbf{w}_k \equiv \mathbf{w}_{[4]\setminus\{k\}}$ and $\alpha_k \equiv \alpha_{[4]\setminus\{k\}}$ (e.g. $\mathbf{u}_1 \equiv \mathbf{u}_{\{2,3,4\}}$).

A. EP Beamformer, $P = 2$

Based on $\mathbf{V}_1$, we have $\mathcal{Z}(1) = \mathcal{Z}(3) = \{W_1 \mid W \in \mathcal{F}\}$, $\mathcal{Z}(2) = \mathcal{Z}(4) = \{W_2 \mid W \in \mathcal{F}\}$. Using EP beamformer (5), the transmission vector is built as

$$\mathbf{x} = \sqrt{P_T/4}\big[A_2\mathbf{u}_3 + B_1\mathbf{u}_4 + C_2\mathbf{u}_1 + D_1\mathbf{u}_2\big] \ . \tag{11}$$

Based on ZF definition, the third term is nulled at user 1; i.e.

$$y_1 = \sqrt{P_T/4}\big[A_2\mathbf{h}_1^T\mathbf{u}_3 + B_1\mathbf{h}_1^T\mathbf{u}_4 + D_1\mathbf{h}_1^T\mathbf{u}_2\big] + z_1 \ . \tag{12}$$

Now user 1 can reconstruct and remove the second and third terms in (12) using $\mathcal{Z}(1)$; and decode A_2 with SINR

$$\text{SINR}_1^1 = \frac{P_T}{4}\frac{|\mathbf{h}_1^T\mathbf{u}_{\{1,2,4\}}|^2}{N_0} \ . \tag{13}$$

Following the same procedure for all users, the symmetric rate can be calculated as

$$r_s^{EP,\mathbf{V}_1} = \frac{P_T}{4N_0}\min \mathcal{L}_{EP}^{\mathbf{V}_1} \ , \tag{14}$$

where $\mathcal{L}_{EP}^{\mathbf{V}_1} = \big\{|\mathbf{h}_1^T\mathbf{u}_3|^2, |\mathbf{h}_2^T\mathbf{u}_4|^2, |\mathbf{h}_3^T\mathbf{u}_1|^2, |\mathbf{h}_4^T\mathbf{u}_2|^2\big\}$. For the total delivery time we have

$$T_{EP}^{\mathbf{V}_1} = \frac{1}{2}\frac{1}{r_s^{EP,\mathbf{V}_1}} = \frac{1}{2}\frac{N_0}{P_T}\frac{4}{\min \mathcal{L}_{EP}^{\mathbf{V}_1}} \ . \tag{15}$$

B. EP Beamformer, $P = 4$

Following the same procedure in $P = 2$, $\mathbf{x}$ is built as

$$\begin{aligned} \mathbf{x} = \sqrt{P_T/4}\big[&(A_2 \oplus C_1)\mathbf{u}_4 + (A_3 \oplus C_4)\mathbf{u}_2 \\ &+(B_3 \oplus D_2)\mathbf{u}_1 + (B_4 \oplus D_1)\mathbf{u}_3\big] \ , \end{aligned} \tag{16}$$

and user 1 receives

$$\begin{aligned} y_1 = \sqrt{P_T/4}\big[&(A_2 \oplus C_1)\mathbf{h}_1^T\mathbf{u}_4 + (A_3 \oplus C_4)\mathbf{h}_1^T\mathbf{u}_2 \\ &+ (B_4 \oplus D_1)\mathbf{h}_1^T\mathbf{u}_3\big] + z_1 \ . \end{aligned} \tag{17}$$

User 1 can reconstruct and remove the third term using $\mathcal{Z}(1)$, and decode $(A_2 \oplus C_1)$ and $(A_3 \oplus C_4)$ with SINR values

$$\text{SINR}_1^1 = \frac{P_T}{4}\frac{|\mathbf{h}_1^T\mathbf{u}_3|^2}{N_0}, \ \text{SINR}_1^2 = \frac{P_T}{4}\frac{|\mathbf{h}_1^T\mathbf{u}_2|^2}{N_0} \ . \tag{18}$$

Next, user 1 again uses $\mathcal{Z}(1)$ to extract its requested terms A_2, A_3. Following the same procedure for all users, we have

$$r_s^{EP,\mathbf{V}_2} = \frac{P_T}{4N_0} \min \mathcal{L}_{EP}^{\mathbf{V}_2} \ , \tag{19}$$

where

$$\begin{aligned} \mathcal{L}_{EP}^{\mathbf{V}_2} = \big\{ &|\mathbf{h}_1^T\mathbf{u}_4|^2, |\mathbf{h}_1^T\mathbf{u}_2|^2, |\mathbf{h}_2^T\mathbf{u}_1|^2, |\mathbf{h}_2^T\mathbf{u}_3|^2, \\ &|\mathbf{h}_3^T\mathbf{u}_4|^2, |\mathbf{h}_3^T\mathbf{u}_2|^2, |\mathbf{h}_4^T\mathbf{u}_3|^2, |\mathbf{h}_4^T\mathbf{u}_1|^2 \big\} \ . \end{aligned} \tag{20}$$

Then for the total delivery time we have

$$T_{EP}^{\mathbf{V}_2} = \frac{1}{4} \frac{1}{r_s^{EP,\mathbf{V}_2}} = \frac{1}{4} \frac{N_0}{P_T} \frac{4}{\min \mathcal{L}_{EP}^{\mathbf{v}_2}} \ . \tag{21}$$

C. EP Beamformer, $P = 6$

Following the same procedure, each user has to decode its requested terms through a MAC channel of size 3; and

$$r_s^{EP,\mathbf{V}_3} = \frac{P_T}{4N_0} \min \mathcal{L}_{EP}^{\mathbf{V}_3} \ , \tag{22}$$

where $\mathcal{L}_{EP}^{\mathbf{V}_3}$ is defined as

$$\begin{aligned} \mathcal{L}_{EP}^{\mathbf{V}_3} = \big\{ &|\mathbf{h}_1^T\mathbf{u}_4|^2, |\mathbf{h}_1^T\mathbf{u}_2|^2, |\mathbf{h}_1^T\mathbf{u}_3|^2, |\mathbf{h}_2^T\mathbf{u}_1|^2, \\ &|\mathbf{h}_2^T\mathbf{u}_3|^2, |\mathbf{h}_2^T\mathbf{u}_4|^2, |\mathbf{h}_3^T\mathbf{u}_4|^2, |\mathbf{h}_3^T\mathbf{u}_2|^2, \\ &|\mathbf{h}_3^T\mathbf{u}_1|^2, |\mathbf{h}_4^T\mathbf{u}_3|^2, |\mathbf{h}_4^T\mathbf{u}_1|^2, |\mathbf{h}_4^T\mathbf{u}_2|^2 \big\} \ . \end{aligned} \tag{23}$$

The total delivery time can then be calculated as

$$T_{EP}^{\mathbf{V}_3} = \frac{1}{6} \frac{1}{r_s^{EP,\mathbf{V}_3}} = \frac{1}{6} \frac{N_0}{P_T} \frac{4}{\min \mathcal{L}_{EP}^{\mathbf{v}_3}} \ . \tag{24}$$

D. PL Beamformer

According to (6), using PL structure instead of EP, the power coefficient $P_T/4$ is replaced with $\alpha_{\mathcal{S}} P_T$. This causes r_s to be a function of $\alpha_{\mathcal{S}}$; which should be optimized in order to maximize r_s. Starting from the case $P = 2$, we define

$$\mathcal{L}_{PL}^{\mathbf{V}_1} = \big\{ \alpha_3 |\mathbf{h}_1^T\mathbf{u}_3|^2, \alpha_4 |\mathbf{h}_2^T\mathbf{u}_4|^2, \alpha_1 |\mathbf{h}_3^T\mathbf{u}_1|^2, \alpha_2 |\mathbf{h}_4^T\mathbf{u}_2|^2 \big\} \ .$$

The symmetric rate is then calculated as

$$r_s^{PL,\mathbf{V}_1} = \frac{P_T}{N_0} \max_{\alpha_{\mathcal{S}}} \min \mathcal{L}_{PL}^{\mathbf{V}_1} \ , \tag{25}$$

and the delivery time will be

$$T_{PL}^{\mathbf{V}_1} = \frac{1}{2} \frac{1}{r_s^{PL,\mathbf{V}_1}} = \frac{1}{2} \frac{N_0}{P_T} \frac{1}{\max_{\alpha_{\mathcal{S}}} \min \mathcal{L}_{PL}^{\mathbf{V}_1}} \ . \tag{26}$$

Delivery times for $P = 4, 6$ are calculated similarly.

E. OB Beamformer

Using OB instead of EP, $P_T/4$ is replaced with $\alpha_{\mathcal{S}} P_T$ and $\mathbf{v}_{\mathcal{S}}$ is used instead of $\mathbf{u}_{\mathcal{S}}$. This causes r_s to be a function of both $\alpha_{\mathcal{S}}$ and $\mathbf{v}_{\mathcal{S}}$; and SINR terms to include interference from unwanted terms. Considering the case $P = 2$, we have

$$\mathbf{x} = \sqrt{P_T} \big[A_2 \sqrt{\alpha_3} \mathbf{v}_3 + B_1 \sqrt{\alpha_4} \mathbf{v}_4 + C_2 \sqrt{\alpha_1} \mathbf{v}_1 + D_1 \sqrt{\alpha_2} \mathbf{v}_2 \big],$$

and user 1 receives

$$\begin{aligned} y_1 = \sqrt{P_T} \big[&A_2 \sqrt{\alpha_3} \mathbf{h}_1^T \mathbf{v}_3 + B_1 \sqrt{\alpha_4} \mathbf{h}_1^T \mathbf{v}_4 \\ &+ C_2 \sqrt{\alpha_1} \mathbf{h}_1^T \mathbf{v}_1 + D_1 \sqrt{\alpha_2} \mathbf{h}_1^T \mathbf{v}_2 \big] + z_1 \ , \end{aligned} \tag{27}$$

from which it can reconstruct and remove the second and fourth terms, using $\mathcal{Z}(1)$. The third term appears as interference however; and hence for decoding A_2 at user 1 we have

$$\text{SINR}_1^1 = \alpha_3 P_T \frac{|\mathbf{h}_1^T \mathbf{v}_3|^2}{|\mathbf{h}_1^T \mathbf{v}_1|^2 + N_0} \ . \tag{28}$$

the symmetric rate can then be calculated as

$$r_s^{OB,\mathbf{V}_1} = \frac{P_T}{N_0} \max_{\alpha_{\mathcal{S}}, \mathbf{v}_{\mathcal{S}}} \min \mathcal{L}_{OB}^{\mathbf{V}_1} \ , \tag{29}$$

where $\mathcal{L}_{OB}^{\mathbf{V}_1}$ is defined as

$$\begin{aligned} \mathcal{L}_{OB}^{\mathbf{V}_1} = \big\{ &\frac{\alpha_3 N_0 |\mathbf{h}_1^T \mathbf{v}_3|^2}{\alpha_1 |\mathbf{h}_1^T \mathbf{v}_1|^2 + N_0}, \frac{\alpha_4 N_0 |\mathbf{h}_2^T \mathbf{v}_4|^2}{\alpha_2 |\mathbf{h}_2^T \mathbf{v}_2|^2 + N_0}, \\ &\frac{\alpha_1 N_0 |\mathbf{h}_3^T \mathbf{v}_1|^2}{\alpha_3 |\mathbf{h}_3^T \mathbf{v}_3|^2 + N_0}, \frac{\alpha_2 N_0 |\mathbf{h}_1^T \mathbf{v}_2|^2}{\alpha_4 |\mathbf{h}_4^T \mathbf{v}_4|^2 + N_0} \big\} \ . \end{aligned} \tag{30}$$

Finally, for delivery time we have

$$T_{OB}^{\mathbf{V}_1} = \frac{1}{2} \frac{1}{r_s^{OB,\mathbf{v}_1}} = \frac{1}{2} \frac{N_0}{P_T} \frac{1}{\max_{\alpha_{\mathcal{S}}, \mathbf{v}_{\mathcal{S}}} \min \mathcal{L}_{OB}^{\mathbf{V}_1}} \ . \tag{31}$$

Delivery times for $P = 4, 6$ are calculated similarly. Defining

$$\begin{aligned} \tilde{\mathcal{S}} &= \{ \mathcal{S} \subseteq [K] \, ; \, |\mathcal{S}| = t + 1 \} \ , \\ \mathcal{S}(k) &= \big\{ \mathcal{S} \in \tilde{\mathcal{S}} \ \big| \ k \in \mathcal{S}; \ \exists p \in \Phi(\mathcal{S}) : \mathbf{V}[p, k] = 0 \big\} \ , \\ \overline{\mathcal{S}}(k) &= \big\{ \mathcal{S} \in \tilde{\mathcal{S}} \ \big| \ k \notin \mathcal{S} \big\} \ , \end{aligned} \tag{32}$$

for a generic network with cache placement matrix $\mathbf{V}$ we have

$$\mathcal{L}_{OB}^{\mathbf{V}} = \left\{ \frac{N_0 \alpha_{\mathcal{S}_i^k} |\mathbf{h}_k^T \mathbf{v}_{\mathcal{S}_i^k}|^2}{\sum\limits_{\mathcal{S}_j^k \in \overline{\mathcal{S}}(k)} \alpha_{\mathcal{S}_j^k} |\mathbf{h}_k^T \mathbf{v}_{\mathcal{S}_j^k}|^2 + N_0}; \begin{matrix} \forall k \in [K] \\ \forall \mathcal{S}_i^k \in \mathcal{S}(k) \end{matrix} \right\}. \tag{33}$$

F. Generic Networks

Consider a generic network with parameters K, L, t, in which $K = t + L$ and cache placement is performed according to placement matrix $\mathbf{V}$, with dimensions $P \times K$. Then, the total delivery time for EP, PL and OB beamformer structures at low-SNR can be approximately calculated as

$$\begin{aligned} T_{EP}^{\mathbf{V}} &= \frac{1}{P} \frac{N_0}{P_T} \frac{n(\mathbf{V})}{\min \mathcal{L}_{PE}^{\mathbf{V}}} \ , \\ T_{PL}^{\mathbf{V}} &= \frac{1}{P} \frac{N_0}{P_T} \frac{1}{\max_{\alpha_{\mathcal{S}}} \min \mathcal{L}_{PL}^{\mathbf{V}}} \ , \\ T_{OB}^{\mathbf{V}} &= \frac{1}{P} \frac{N_0}{P_T} \frac{1}{\max_{\alpha_{\mathcal{S}}, \mathbf{v}_{\mathcal{S}}} \min \mathcal{L}_{OB}^{\mathbf{V}}} \ , \end{aligned} \tag{34}$$

where $n(\mathbf{V})$ is the number of $\mathcal{S}$ sets for which $\Phi(\mathcal{S}) \neq \varnothing$; and for any $\gamma \in \{\text{EP}, \text{PL}, \text{OB}\}$, $|\mathcal{L}_\gamma^{\mathbf{V}}| = K \times m(\mathbf{V}) = PL$.

IV. Numerical Results and Discussion

According to (34), delivery time $T_\gamma^{\mathbf{V}}$ is proportional to $\frac{1}{P}$; i.e increasing P decreases $T_\gamma^{\mathbf{V}}$. In fact, increasing P enables data to be delivered in smaller chunks in parallel (over the MAC channel); which is characterized as *efficiency index* in [10]. However, increasing P also increases $|\mathcal{L}_\gamma^{\mathbf{V}}|$, resulting in $\min \mathcal{L}_\gamma^{\mathbf{V}}$ to become smaller and $T_\gamma^{\mathbf{V}}$ to be increased.

On the other hand, beamformer structure γ only affects $T_\gamma^{\mathbf{V}}$ through $\mathcal{L}_\gamma^{\mathbf{V}}$. For $\gamma = \text{EP}$, the value of $\mathcal{L}_{EP}^{\mathbf{V}}$ is deterministic. However, if $\gamma \neq \text{PL}$, $\min \mathcal{L}_\gamma^{\mathbf{V}}$ is a function of $\alpha_{\mathcal{S}}$ (and $\mathbf{v}_{\mathcal{S}}$, if $\gamma = \text{OB}$). So $T_\gamma^{\mathbf{V}}$ is calculated through an optimization problem, with $n(\mathbf{V})$ variables $\alpha_{\mathcal{S}}$ and one constraint $\sum_{\mathcal{S}} \alpha_{\mathcal{S}} = 1$ for $\gamma = \text{PL}$; and $(L+1)n(\mathbf{V})$ variables ($\mathbf{v}_{\mathcal{S}} \in \mathbb{C}^L$) and $1 + n(\mathbf{V})$ total constraints ($\|\mathbf{v}_{\mathcal{S}}\| = 1$), for $\gamma = \text{OB}$.

To compare the system performance for various P and γ selections, we use the cumulative distribution function (CDF)

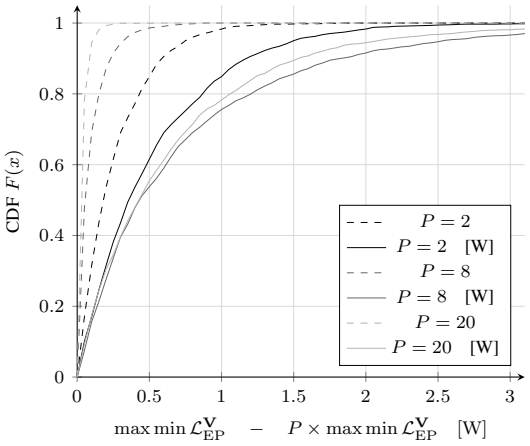

Fig. 4: Empirical CDF, $\gamma = \mathrm{EP}$

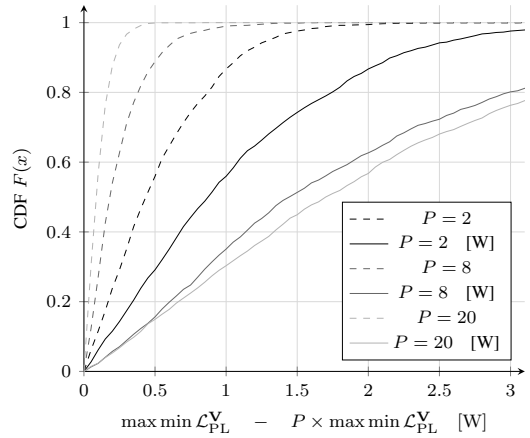

Fig. 5: Empirical CDF, $\gamma = \mathrm{PL}$

of $\max\min \mathcal{L}^{\mathbf{V}}_{\gamma}$, denoted by $F(\mathcal{L}^{\mathbf{V}}_{\gamma})$ for simplicity. As the terms inside $\mathcal{L}^{\mathbf{V}}_{\gamma}$ are correlated to each other, it is difficult to find a closed-form expression for $F(\mathcal{L}^{\mathbf{V}}_{\gamma})$; and so we proceed with a numerical approach. In Figures 4-6 we have plotted the empirical results for $F(\mathcal{L}^{\mathbf{V}}_{\gamma})$ and $P \times F(\mathcal{L}^{\mathbf{V}}_{\gamma})$, for $\gamma \in \{\mathrm{EP}, \mathrm{PL}, \mathrm{OB}\}$ and $P \in \{2, 8, 20\}$. Network parameters are set to $K = 6$, $t = 3$, $L = 3$.

According to the figures, regardless of γ, the value of $F(\mathcal{L}^{\mathbf{V}}_{\gamma})$ decreases as we increase P. This is due to the fact that increasing P increases $|\mathcal{L}^{\mathbf{V}}_{\gamma}|$; i.e. the minimum is taken over a larger number of random variables, resulting in higher probability for a smaller result. However, comparing $P \times F(\mathcal{L}^{\mathbf{V}}_{\gamma})$ reveals that the decrement in $F(\mathcal{L}^{\mathbf{V}}_{\gamma})$ is dependent on γ indeed. In fact, for $\gamma = \mathrm{EP}$, the decrement in $F(\mathcal{L}^{\mathbf{V}}_{\gamma})$ from $P = 8$ to $P = 20$ is so large that even multiplication by P is not enough for its compensation. However, for $\gamma = \mathrm{PL}$ the decrement is almost compensated after the multiplication; and for $\gamma = \mathrm{OB}$ the compensation is so large that $F(\mathcal{L}^{\mathbf{V}}_{\gamma})$ is improved by increasing P. This is in line with the rate behavior with respect to P and γ, as reviewed in Section II-E[1].

[1]As mentioned in Section III, the low-SNR approximation has a more positive effect on R as P becomes larger. So doing the analysis without the approximation causes more destruction in rate for larger P.

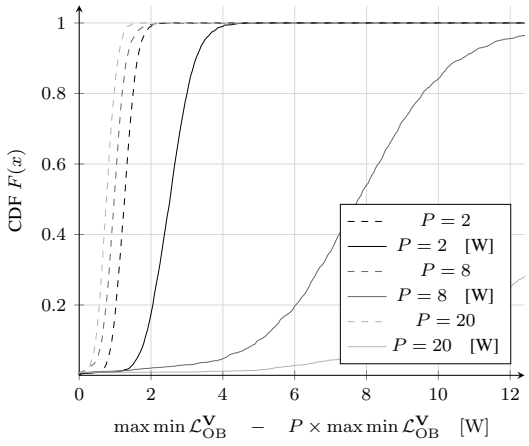

Fig. 6: Empirical CDF, $\gamma = \mathrm{OB}$

V. Conclusion and Future Work

We studied the joint effect of subpacketization P and beamformer structure γ on the rate performance of multi-antenna Coded Caching (CC) setups. Using a low-SNR approximation, we provided simple closed-form rate expressions, and used numerical simulations for performance comparison of various schemes. The results indicate that P and γ jointly affect the rate; i.e. based on γ, P might improve or deteriorate the achievable rate. The results are limited to a specific class of networks with $K = t+L$. Removing this constraint and taking a more theoretical approach (compared to the numerical one in this paper), are parts of the ongoing research.

References

[1] V. N. I. Cisco, "Cisco visual networking index: Forecast and trends, 2017–2022," *White Paper*, vol. 1, 2018.

[2] 6Genesis, "Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence," *White Paper*, vol. 1, 2019.

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

[4] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2792–2807, 2018.

[5] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical coded caching," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3212–3229, 2016.

[6] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Transactions on Networking (TON)*, vol. 24, no. 2, pp. 836–845, 2016.

[7] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1176–1188, 2018.

[8] M. Salehi, A. Tölli, and S. P. Shariatpanahi, "A Multi-Antenna Coded Caching Scheme with Linear Subpacketization," in *2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[9] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multicast beamformer design for coded caching," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 1914–1918.

[10] M. Salehi, A. Tölli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization-Rate Trade-off in Multi-Antenna Coded Caching," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[11] 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.

[12] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multi-antenna Interference Management for Coded Caching," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.