Multi-Antenna Coded Caching Analysis in Finite SNR and Finite Subpacketization

Hui Zhao*, Eleftherios Lampiris*, Giuseppe Caire†, and Petros Elia*
*Communication Systems Department, EURECOM, 06410 Sophia Antipolis, France
†Electrical Engineering and Computer Science Department, Technical University of Berlin, 10587 Berlin, Germany
Email: hui.zhao@eurecom.fr; eleftherios.lampiris@eurecom.fr; caire@tu-berlin.de; elia@eurecom.fr;

***Abstract*— We investigate the performance of multi-antenna coded caching delivery algorithms operating under practical constraints. Specifically, under the constraints of finite subpacketization and finite signal-to-noise ratio (SNR) we compare two popular algorithms, one exhibiting low subpacketization while the other transmits multicast messages, hence requiring fewer symbols for communication and resulting in increased power per transmitted symbol. Our work is motivated by the fact that, while many multi-antenna coded caching algorithms exhibit the same asymptotic performance, when factoring practical limitations such as finite subpacketization and finite SNR, the real-world performances differ significantly.**

***Index Terms*—Coded Caching, finite subpacketization, finite SNR, and multiple antennas.**

I. Introduction

Coded caching is a novel caching method, introduced by Maddah-Ali and Niesen [1], which uses the content stored at the end users as side-information, in order to reduce the interference. The main idea in coded caching is that users store parts from some popular files, in such a manner that allows for some partial overlap between the contents of any two caches. Then, during the delivery phase, the transmitter combines together parts of the desired files, by performing XOR additions, and multicasts the resulting message hence, serving multiple users simultaneously. Specifically, in the model considered in [1] a server, who has access to the whole library, is connected via an *error-free shared-link* to K users, each of whom has the capacity to store fraction $\gamma \in [0, 1]$ of the total library content. In the end, the algorithm in [1] manages to simultaneously serve $K\gamma + 1$ users.

A. Multi-Antenna Coded Caching

Another way to boost communications has been the use of multi-antenna arrays. Hence, many works have sought to design algorithms that can exploit the benefits of both coded caching and multiple antennas.

First such efforts created "separable" algorithms, i.e. algorithms which generate multicast messages, as in [1], and then find ways to efficiently communicate these messages to the users via the multiple antennas. For example, the work in [2] proposed multiple antenna algorithms using no or low channel state information (CSI) at the transmitter which relied on the diversity offered from the multiple antennas in order to increase the communication rate. On the other hand, the work in [3], as well as the literature that was inspired by it, focused on the design of beamforming vectors that maximize the minimum (among the users served) rate of a single multicast message.

A different direction, though, looked into combining the two resources in order to increase the Degrees of Freedom (DoF) performance of the system. The first such effort, proposed in [4], merged cache-aided decoding with multi-antenna precoding. In the end, for a system with K users, each having a cache of normalized size γ and which are served via an L-antenna transmitter, the DoF performance achieved by the algorithm in [4] was shown to be

$$\mathcal{D}_L = K\gamma + L. \tag{1}$$

This hinted towards an additive relationship between the two resources. Later, [5] showed that the performance in (1) is exactly optimal under uncoded placement and one-shot linear precoding.

Further, the multi-antenna algorithms proposed in [5]–[13] have produced different improvements, such as lower CSI requirements [5] or the ability to communicate to both cache-aided and cache-less users without loss in performance [11], to name a few.

Although it is shown that in the very high SNR region, the performance of the aforementioned algorithms is the same, this similarity vanishes in low-to-moderate values of the SNR, where the majority of communication systems operate [14]–[16]. Various works have sought to capture the finite SNR performances of different multi-antenna coded caching algorithms [17]–[20]

B. Subpacketization Bottleneck of Coded Caching

While, in theory, coded caching can achieve a DoF that scales linearly with the number of users, to do so it requires the size of each file to be scaling exponentially or near-exponentially with the number of users [21]. Specifically, in order for the full gain of $K\gamma+1$ to be achieved, the algorithm of [1] requires that the number of different subpackets, F_{MAN}, of a file need to be $F_{\text{MAN}} = \binom{K}{K\gamma} \approx (1/\gamma)^{K\gamma}$. This creates a bottleneck, which has been termed the "subpacketization constraint", since even with a modest number of users $K \geq 50$ and cache of normalized size $\gamma \leq 0.1$, the required size of a file needs to be at least 2 GB, when considering the minimum subpacket size to be in the order of 1KB in magnitude. The

This work is supported by the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant agreement no. 725929. (ERC project DUALITY) and by the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant 789190 - CARENET

subpacketization constraint has sparked a research direction which seeks to design algorithms that require small subpacketization, while exhibiting high multicasting gain [21]–[27].

A simpler algorithm compared to those in [22]–[27] has been proposed in [21]. For a target subpacketization F , the main idea in the algorithm in [21] is to encode among $\Lambda < K$ users, such that

$$\Lambda = \arg\max_{K' \leq K} \left\{ \binom{K'}{K'\gamma} \leq F \right\} \tag{2}$$

with the corresponding DoF being $\Lambda\gamma + 1$ which is also referred to as the *nominal coded caching gain* [28].

The limiting gains of coded caching under finite subpacketization have been further constrained in the earlier multi-antenna coded caching algorithms. As an example, works such as [4], [6], [7] achieved the full caching gain as well as the full multiplexing gain, but required the subpacketization to be exponential in both K and L, as we can see from (3)

$$F_{\text{MS}} = \binom{K - K\gamma - 1}{L - 1} \binom{K}{K\gamma} \approx \left(\frac{K}{L}\right)^{L} \left(\frac{1}{\gamma}\right)^{K\gamma}. \tag{3}$$

A radical approach has been proposed in [29], where it was shown that the DoF of $K\gamma + 1$ can be achieved without the "multi-antenna subpacketization" term in (3), and at the same time the subpacketization of the "caching" term can be reduced to approximately the L-th root of the subpacketization of the single antenna case, i.e., $F_{\text{LS}} = \binom{K/L}{K\gamma/L} \approx \sqrt[L]{\binom{K}{K\gamma}}$. Alternatively, we can see that for some target subpacketization F, one can achieve L times the DoF of the single antenna case. This low-subpacketization scheme in [29] is henceforth referred to as the *LSP scheme*.

C. Contributions

The objective of our work is to compare the performance of two different multi-antenna coded caching schemes, each exhibiting its own desirable characteristics, and which schemes will operate under two practical constraints i) a finite subpacketization constraint, and ii) finite SNRs. As discussed above, subpacketization is a limiting factor, and no known single-antenna coded caching algorithm has been able to exhibit a gain of more than 5, for reasonable values of the practical parameters (e.g., the maximum file size and the minimum packet size). Further, finite SNR analysis allows us to evaluate the performance of the considered coded caching algorithms in practical SNR regimes. While these two analysis have been performed individually, i.e., finite SNR analysis with $F \to \infty$ and finite subpacketization analysis with SNR $\to \infty$, our work here presents the first effort to understand the performance of these algorithms under both constraints.

The first algorithm we will consider is the low-subpacketization scheme — LSP in [29], and we will compare it with the performance of a variation of the multi-server algorithm in [11], which retains all the desirable properties of the original multi-server algorithm while reducing complexity. We will refer to the multi-server variant as the *MSV scheme*. The first difference between the two schemes is the number of users that they can simultaneously serve for some subpacketization level F. Speifically, the maximum number of users served at a time under the considered two schemes are respectively,

$$\text{DoF}_{\text{LSP}} = L \cdot (\Lambda\gamma + 1), \quad \text{DoF}_{\text{MSV}} = \Lambda\gamma + L, \tag{4}$$

where Λ is determined by (2). That is, the LSP scheme provides higher DoF than the MSV scheme.

A second difference is that the LSP scheme in [29] communicates one subfile per symbol, *i.e., coded caching with unicasting transmissions*, while the MSV scheme in [11] sends $\Lambda\gamma + L$ subfiles using $\frac{L+\Lambda\gamma}{1+\Lambda\gamma}$ symbols, i.e. each symbol carries, on average, $\Lambda\gamma + 1$ subfiles. In other words, the MSV scheme in [11] utilizes *multicasting multi-group transmissions*, thereby making each symbol more efficient.

We design two-layered linear precoders separately for the physical-layer (PLY) transmissions in the considered LSP and MSV schemes, where the outer-layer precoding is used for interference cancellation, and beamforming is applied in the inner-layer precoding. We evaluate the performance of each algorithm through numerical simulations, while we use as a benchmark the uncoded caching scenario.

Numerical results show that the LSP scheme always provides the largest overall throughput in finite SNRs. In contrast, there exist cut-off points in the SNR region for which uncoded caching can outperform MSV. Specifically, large-antenna arrays and low SNR values enable the uncoded caching delivery to achieve higher throughput compared to MSV.

Notation: We use $\mathbb{C}$ to denote the set of complex numbers. $\mathbf{0}_L \in \mathbb{C}^{L\times 1}$ denotes the vector with all zero elements, and $\mathbf{I}_L \in \mathbb{C}^{L\times L}$ represents the identity matrix. $\mathcal{CN}(\mathbf{0}_L, \mathbf{I}_L)$ denotes the multivariate Gaussian distribution with mean vector $\mathbf{0}_L$ and covariance matrix $\mathbf{I}_L$. For a positive integer n, we use $[n]$ to define the set $[n] \triangleq \{1, 2, \cdots, n\}$, and use $|\cdot|$ to denote the cardinality of a set or the magnitude of a complex number. $||\cdot||$ stands for the norm-2 operator of a vector. The suberscripts T and H denote the non-conjugate and conjugate transpose of a matrix, respectively. $\text{Tr}\{\cdot\}$ denotes the trace operator of a matrix. The operator $\lfloor\cdot\rfloor$ denotes the nearest integer less than or equal to the argument.

II. System Model and delivery algorithms

We consider a system with a base station (BS) equipped with L antennas, which has access to a library $\mathcal{F}$ of N popular files, each of which is composed of F indivisible data units, henceforth called subfiles. The BS serves K users, each endowed with a cache able to store $M \cdot F$ data units, where $M < N$ or equivalently each user can store fraction $\gamma \triangleq \frac{M}{N}$ of the total library. In this paper, we separately consider the LSP and the MSV coded caching schemes to boost the delivery performance. Both of the two schemes consist of two dedicated phases — the placement phase and the delivery phase. In the following, we will elaborate those two schemes.

Placement Phase: The placement of content at the users uses the algorithm of [21]. Specifically, one selects variable Λ according to (2) and continues with the design of Λ different caches, as in the algorithm of [1] and randomly assigns a cache to every user. In other words, variable Λ signifies the number of caches whose content partially overlaps and, hence,

for every cache there will be approximately $\frac{K}{\Lambda}$ users who will receive that cache.

More formally, during the placement phase each file is divided as $W_n \longrightarrow \left\{W_n^{\mathcal{T}}, \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\right\}$, and each different cache $g \in [\Lambda]$ is populated as $\mathcal{Z}_g = \left\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma, \mathcal{T} \ni g, \forall n \in [N]\right\}$. By randomly assigning one of the Λ caches to each user we have that approximately $\frac{K}{\Lambda}$ users have access to the exact same content. Henceforth, we will refer to users who have the same cache as *belonging to the same (shared-cache) group*.

Delivery phase: The delivery phase begins with every user requesting simultaneously a file. For simplicity in the analysis we assume that each user requests a different file. We use $d_{g,b} \in [N]$ to denote the file index intended by $\mathrm{U}_{g,b}$ — the b-th user in group g for some $b \in [\frac{K}{\Lambda}]$ and $g \in [\Lambda]$.

1) The MSV Scheme [20]: The first delivery algorithm is based on the Multi-Server work in [20]. The main idea is to construct multicast messages, by performing XOR operations on subfiles, as in the algorithm of [1], and transmitting each multicast message by multiplying it with a precoder designed to nullify the interference caused at the non-intended users. Each multicast message can be "steered-away" from at most $L-1$ users, and since each multicast message can serve a maximum of $\Lambda\gamma+1$ users, it means that the maximum number of simultaneously served users by the MSV scheme is $\Lambda\gamma+L$.

For some subset of the groups $\Psi \subseteq [\Lambda]$, $|\Psi| = \Lambda\gamma+1$, and some $v \in \left[\frac{K}{\Lambda}\right]$ signifying the v-th user of a group, a multicast message takes the form

$$X_{\Psi,v} = \chi\left(\left\{W_{d_{g,v}}^{\Psi\setminus\{g\}}\right\}_{g\in\Psi}\right), \tag{5}$$

where $\chi(\cdot)$ denotes the adopted transmission scheme. Then, by selecting a set of $\lfloor\frac{L+\Lambda\gamma}{1+\Lambda\gamma}\rfloor$ multicast messages with non-overlapping users one can communicate them using multi-antenna precoding.

2) Delivery Phase of the LSP Scheme [29]: The second delivery algorithm is based on the method introduced in [29] and which can achieve multiplicative gains. For some $J \leq L$, one can serve using the LSP scheme $J \cdot (\Lambda\gamma+1)$ users simultaneously. We continue with a high level description of the scheme, and we refer the interested reader to [29] for the complete description.

At the beginning of a time-slot a set of $\Lambda\gamma+1$ groups are selected and then a set of J users are selected from each group. For each set of users belonging to the same group a vector of length J is formed and multiplied by a precoder designed to nullify the intra-group interference. Then, all the $\Lambda\gamma+1$ vectors are added together and transmitted.

III. Linear Precoding Analysis

In this section, we focus separately on a specific transmission stage to serve a group-set Ψ in the MSV scheme and a group-set Φ in the LSP scheme by considering linear precoders during the delivery phase.

We consider Rayleigh fading to model the PLY wireless channel. The channel gain vector from the BS to $\mathrm{U}_{g,b}$ is $\mathbf{h}_{g,b} = \beta_{g,b}\mathbf{n}_{g,b} \in \mathbb{C}^{L\times 1}$ where $\beta_{g,v}$ accounts for the large-scale fading and/or path-loss, and $\mathbf{n}_{g,b} \sim \mathcal{CN}(\mathbf{0}_L, \mathbf{I}_L)$ is a standard complex Gaussian random vector. To facilitate the analysis, we assume that K *approaches to infinity, while* Λ *remains constant, which implies that* $B = \frac{K}{\Lambda} \to \infty$. This assumption on $B \to \infty$ allows us to use the aggregated coded caching (ACC) idea in [15] such that summing the rates in each served user is the overall throughput without considering the limitation from the user with the weakest signal link, i.e., removing the worst-user bottleneck [28].

A. Linear Precoding in MSV Scheme

As described in the delivery phase of the MSV scheme in Section II, for the service to a specific group-set Ψ with V active users for receiving signals in each group, let $\mathbf{s}_\Psi = [s_1, s_2, \cdots, s_V]^T \in \mathbb{C}^{V\times 1}$ be the data vector for the V multicasting groups[1], where s_v is intended by the v-th multicasting group. The transmitted signal is designed as

$$\mathbf{x}_\Psi = \mathbf{M}\mathbf{s}_\Psi \in \mathbb{C}^{L\times 1}, \tag{6}$$

where $\mathbf{M} \in \mathbb{C}^{L\times V}$ denotes the linear precoding matrix adopted at the BS and its v-th column is denoted by $\mathbf{m}_v \in \mathbb{C}^{L\times 1}$. We assume that P_t is the maximum transmit power at the BS, and then We have that

$$\mathbb{E}\{||\mathbf{x}_\Psi||^2\} = \mathrm{Tr}\left\{\mathbf{M}\mathbf{M}^H\right\} \leq P_t. \tag{7}$$

The received signal at $\mathrm{U}_{g,v}$ is

$$\begin{aligned} y_{g,v} &= \mathbf{h}_{g,v}^T\mathbf{x}_\Psi + z_{g,v} = \mathbf{h}_{g,v}^T\mathbf{M}\mathbf{s}_\Psi + z_{g,v} \\ &= \mathbf{h}_{g,v}^T\mathbf{m}_v s_v + \sum\nolimits_{\ell=1,\ell\neq v}^{V} \mathbf{h}_{g,v}^T\mathbf{m}_\ell s_\ell + z_{g,v}, \end{aligned} \tag{8}$$

where $z_{g,v}$ is the Additive White Gaussian Noise (AWGN) with power N_0. The corresponding signal-to-interference plus noise ratio (SINR) is

$$\mathrm{SINR}_{g,v} = \frac{|\mathbf{h}_{g,v}^T\mathbf{m}_v|^2}{N_0 + \sum_{\ell=1,\ell\neq v}^{V}|\mathbf{h}_{g,v}^T\mathbf{m}_\ell|^2}. \tag{9}$$

By employing the ACC idea [15], the BS can achieve multi-rate transmissions to different users within a multicasting group. When a user obtains its intended subfile, another user having the same cached content replaces it without delay, which keeps the multi-rate transmissions. Thus, the sum-rate to the $(\Lambda\gamma+1)V$ users served in a time can be written as

$$R_{\mathrm{sum}}^{\mathrm{MSV}} = \sum_{g\in\Psi, v\in[V]} \log_2\left(1 + \frac{|\mathbf{h}_{g,v}^T\mathbf{m}_v|^2}{N_0 + \sum\limits_{\ell\in[V],\ell\neq v}|\mathbf{h}_{g,v}^T\mathbf{m}_\ell|^2}\right). \tag{10}$$

B. Linear Precoding in LSP Scheme

According to the description for the delivery phase in the LSP scheme in Section II, for a specific group-set Φ with J users selected for service in each group, let $\mathbf{s}_\Phi \triangleq \{s_{q,j}\}_{q\in\Phi, j\in[J]} \in \mathbb{C}^{(\Lambda\gamma+1)J\times 1}$ be the unicasting data vector, where $s_{q,j}$ is the transmitted data symbol to $\mathrm{U}_{q,j}$. Let $\mathbf{r}_{q,j} \in \mathbb{C}^{L\times 1}$ be the precoding vector to $\mathrm{U}_{q,j}$. The transmitted signal after linear precoding is

$$\mathbf{x}_\Phi = \mathbf{R}\mathbf{s}_\Phi \in \mathbb{C}^{L\times 1}, \tag{11}$$

where $\mathbf{R} \triangleq \{\mathbf{r}_{q,j}\}_{q\in\Phi, j\in[J]} \in \mathbb{C}^{L\times(\Lambda\gamma+1)J}$ is the linear precoding matrix, which satisfies that $\mathrm{Tr}\left\{\mathbf{R}\mathbf{R}^H\right\} \leq P_t$.

[1]The multicasting group should not be confused by the shared-cache group.

ISBN 978-3-8007-5686-5

The received signal at $\mathrm{U}_{q,j}$ is given by

$$y_{q,j} = \mathbf{h}_{q,j}^T \mathbf{r}_{q,j} s_{q,j} + \underbrace{\sum_{j'=1, j' \neq j}^{J} \mathbf{h}_{q,j}^T \mathbf{r}_{q,j'} s_{q,j'}}_{\text{intra-group interference}} + \underbrace{\sum_{q' \in \Psi, q' \neq q} \sum_{j'=1}^{J} \mathbf{h}_{q,j}^T \mathbf{r}_{q',j'} s_{q',j'}}_{\text{inter-group interference}} + z_{q,j}. \quad (12)$$

After removing the inter-group interference, the corresponding SINR for information decoding at $\mathrm{U}_{q,j}$ is

$$\mathrm{SINR}_{q,j} = \frac{\left|\mathbf{h}_{q,j}^T \mathbf{r}_{q,j}\right|^2}{N_0 + \sum_{j'=1, j' \neq j}^{J} \left|\mathbf{h}_{q,j}^T \mathbf{r}_{q,j'}\right|^2} \quad (13)$$

By applying the ACC idea [15] in the LSP scheme, when the transmission to a user is finished, another user from the same shared-cache group replaces it for service without delay. This allows us to keep the multi-rate transmission and get out the worst-group constraint. The sum-rate over the simultaneously served $(\Lambda\gamma + 1)J$ users is written as

$$R_{\mathrm{sum}}^{\mathrm{LSP}} = \sum_{q \in \Phi, j \in [J]} \log_2 \left(1 + \frac{\left|\mathbf{h}_{q,j}^T \mathbf{r}_{q,j}\right|^2}{N_0 + \sum\limits_{j' \in [J], j' \neq j} \left|\mathbf{h}_{q,j}^T \mathbf{r}_{q,j'}\right|^2}\right). \quad (14)$$

IV. Linear Precoding Optimization

In this section, we describe the linear precoder design seperately for the MSV and the LSP schemes. We consider the max-min fairness (MMF) to optimize the precoder, where the minimum SINR among the simultaneously served users is maximized. By referring to (9) and (13), the MMF optimization problems in the MSV and the LSP schemes are formulated respectively as

$$\mathbf{M}^* = \underset{\mathbf{M} \in \mathbb{C}^{L \times V}}{\arg\max} \min_{g \in \Psi, v \in [V]} \frac{|\mathbf{h}_{g,v}^T \mathbf{m}_v|^2}{N_0 + \sum_{\ell=1, \ell \neq v}^{V} |\mathbf{h}_{g,v}^T \mathbf{m}_\ell|^2} \quad \text{s.t. } \mathrm{Tr}\left\{\mathbf{M}\mathbf{M}^H\right\} \leq P_t, \quad (15)$$

$$\mathbf{R}^* = \underset{\mathbf{R} \in \mathbb{C}^{L \times (\Lambda\gamma+1)J}}{\arg\max} \min_{q \in \Phi, j \in [J]} \frac{|\mathbf{h}_{q,j}^T \mathbf{r}_{q,j}|^2}{N_0 + \sum_{j'=1, j' \neq j}^{J} |\mathbf{h}_{q,j}^T \mathbf{r}_{q,j'}|^2} \quad \text{s.t. } \mathrm{Tr}\left\{\mathbf{R}\mathbf{R}^H\right\} \leq P_t, \quad (16)$$

which are NP-hard [30]–[32]. To simplify the design, we employ sub-optimal designs which have been proposed in the MIMO and multi-group multicasting literature [31]. The main idea behind the design is a two-layered precoding algorithm, where the interference is cancelled in the outer-layer, and beamforming is used in the inner-layer.

A. Outer-Layer: Interference Cancellation

1) Outer-Layer in MSV Scheme: Define $m \triangleq L - (\Lambda\gamma + 1)(V - 1)$ and define $\mathbf{H}_{\Psi,-v} \in \mathbb{C}^{\Lambda\gamma V \times L}$ after removing the rows from $(v-1)(\Lambda\gamma+1)+1$ to $v(\Lambda\gamma+1)$ of $\mathbf{H}_\Psi$, where $\mathbf{H}_\Psi \in \mathbb{C}^{(\Lambda\gamma+1)V \times L}$ is the channel matrix from the BS to the selected users in the group-set Ψ. Note that $\mathbf{H}_{\Psi,-v}$ is actually the channel matrix from the BS to all the simultaneously served users in Ψ except the v-th multicasting group. Using QR-based decomposition approach, we can easily obtain an orthogonal basis of the null space of $\mathbf{H}_{\Psi,-v}^H$, which constitutes the matrix $\mathbf{Q}_{\Psi,-v}^{\mathrm{null}} \in \mathbb{C}^{L \times m}$. Therefore, we have that $\mathbf{H}_{\Psi,-v}\mathbf{Q}_{\Psi,-v}^{\mathrm{null}} = \mathbf{0} \in \mathbb{C}^{(|\mathcal{G}|-1)V \times m}$ and $(\mathbf{Q}_{\Psi,-v}^{\mathrm{null}})^H \mathbf{Q}_{\Psi,-v}^{\mathrm{null}} = \mathbf{I}_m$. We use $\mathbf{Q}_{\Psi,-v}^{\mathrm{null}}$ as the outer-layer precoding matrix for interference cancellation in (9). We can write the precoding vector $\mathbf{m}_v$ for the v-th multicasting group as

$$\mathbf{m}_v = \mathbf{Q}_{\Psi,-v}^{\mathrm{null}} \mathbf{c}_v, \text{ for } v \in [V], \quad (17)$$

where $\mathbf{c}_v \in \mathbb{C}^{m \times 1}$ is the inner-layer beamforming vector for the v-th multicasting group. *Note that V should be less than or equal to $\lfloor \frac{\Lambda\gamma+L}{\Lambda\gamma+1} \rfloor$ to guarantee non-empty null space of $\mathbf{H}_{\Psi,-v}^H$.* It is easy to check that

$$||\mathbf{m}_v||^2 = \mathbf{c}_v^H (\mathbf{Q}_{\Psi,-v}^{\mathrm{null}})^H \mathbf{Q}_{\Psi,-v}^{\mathrm{null}} \mathbf{c}_v = ||\mathbf{c}_v||^2. \quad (18)$$

2) Outer-Layer in LSP Scheme: Similarly, define $\eta = L - (J-1)$, and we use $\mathbf{H}_{\Phi,q,-j} \in \mathbb{C}^{(J-1) \times L}$ to denote $\mathbf{H}_{\Phi,q} \triangleq [\mathbf{h}_{q,1}, \cdots, \mathbf{h}_{q,J}]^T$ with the j-th row removed. In the following, we will omit the subscript Φ for notational simplification. Using QR-decomposition, we can obtain the matrix $\mathbf{\Theta}_{q,-j}^{\mathrm{null}} \in \mathbb{C}^{L \times \eta}$ whose columns form an orthogonal basis of the null space of $\mathbf{H}_{q,-j}^H$, and we have that $\mathbf{H}_{q,-j}\mathbf{\Theta}_{q,-j}^{\mathrm{null}} = \mathbf{0} \in \mathbb{C}^{(J-1) \times \eta}$ and $(\mathbf{\Theta}_{q,-j}^{\mathrm{null}})^H \mathbf{\Theta}_{q,-j}^{\mathrm{null}} = \mathbf{I}_\eta$. As we only need to cancel the interference from other users with the same cached content as $\mathrm{U}_{q,j}$, we normally have more freedom (dimensions) to find a beamforming vector that maximizes $\mathrm{SINR}_{q,j}$ in (13). The precoding vector to $\mathrm{U}_{q,j}$ can be written as

$$\mathbf{r}_{q,j} = \mathbf{\Theta}_{q,-j}^{\mathrm{null}} \mathbf{u}_{q,j}, \text{ for } q \in \Phi, j \in [J], \quad (19)$$

where $\mathbf{u}_{q,j} \in \mathbb{C}^{\eta \times 1}$ denotes the inner-layer beamforming vector to $\mathrm{U}_{q,j}$. We also have that $||\mathbf{r}_{q,j}||^2 = ||\mathbf{u}_{q,j}||^2$. *Note that J should be less than or equal to L to guarantee non-empty null-space of $\mathbf{H}_{q,-j}^H$.*

B. Inner-Layer: Beamforming

We now proceed to find optimal beamforming vectors of $\mathbf{c}_v$ in (17) and $\mathbf{u}_{q,j}$ in (19) respectively. By considering the analysis in Section IV-A, it is easy to derive the following optimization problems

$$\mathbf{c}^* = \underset{\mathbf{c} \in \mathbb{C}^{mV}}{\arg\max} \min_{g \in \Phi, v \in [V]} \left|\mathbf{h}_{g,v}^T \mathbf{Q}_{-v}^{\mathrm{null}} \mathbf{c}_v\right|^2, \text{ s.t. } ||\mathbf{c}||^2 \leq P_t, \quad (20)$$

$$\mathbf{u}^* = \underset{\mathbf{u} \in \mathbb{C}^{\eta J(\Lambda\gamma+1)}}{\arg\max} \min_{q \in \Phi, j \in [J]} \left|\mathbf{h}_{q,j}^T \mathbf{\Theta}_{q,-j}^{\mathrm{null}} \mathbf{u}_{q,j}\right|^2, \text{ s.t. } ||\mathbf{u}||^2 \leq P_t, \quad (21)$$

where $\mathbf{c} \triangleq \{\mathbf{c}_v\}_{v=1}^V$ and $\mathbf{u} \triangleq \{\mathbf{u}_{q,j}\}_{q \in \Phi, j \in [J]}$. The optimization problems in (20) and (21) can be easily solved using some software packages, such as in MATLAB and its built-in function "fminimax".

V. Numerical Results

From (4), we know that the theoretical multiplicative performance boost of the LSP compared to the MSV should be $\frac{L(\Lambda\gamma+1)}{\Lambda\gamma+L}$ achieved in the high-SNR limit, while we expect this to be smaller for lower values of the SNR due to the more efficient use of the power. To explore the behavior of the gap at finite SNR values we proceed with the presentation

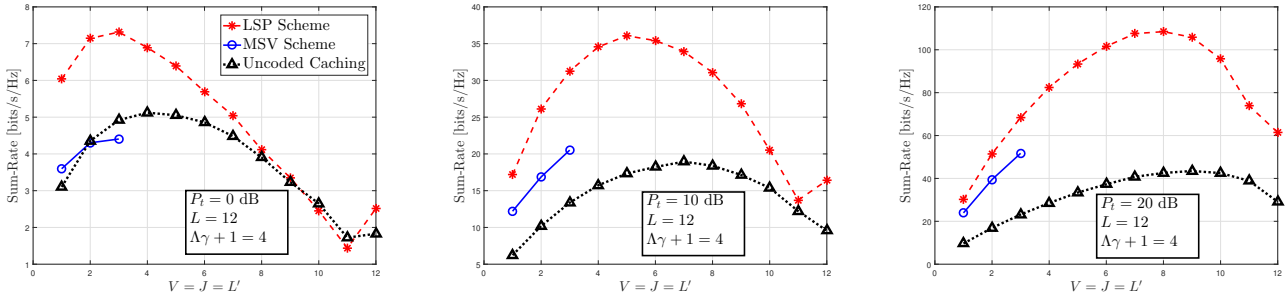

Fig. 1: Sum-rate comparisons for $P_t = 0, 10, 20$ dB, where $L = 12$ and $\Lambda\gamma + 1 = 4$

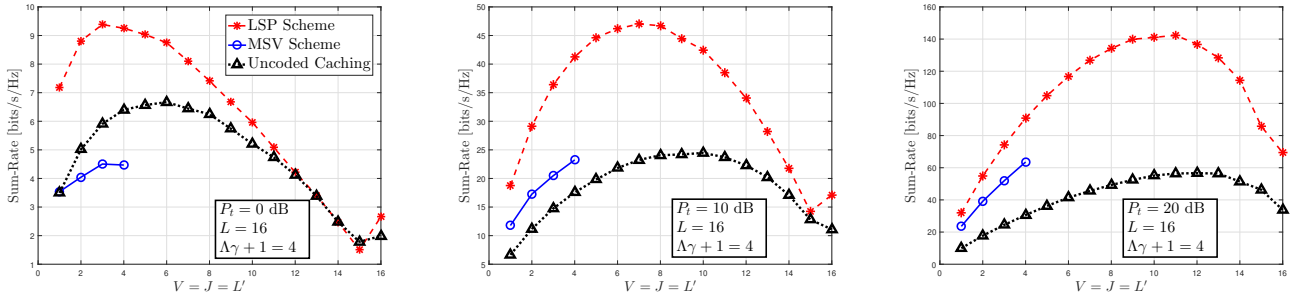

Fig. 2: Sum-rate comparisons for $P_t = 0, 10, 20$ dB, where $L = 16$ and $\Lambda\gamma + 1 = 4$

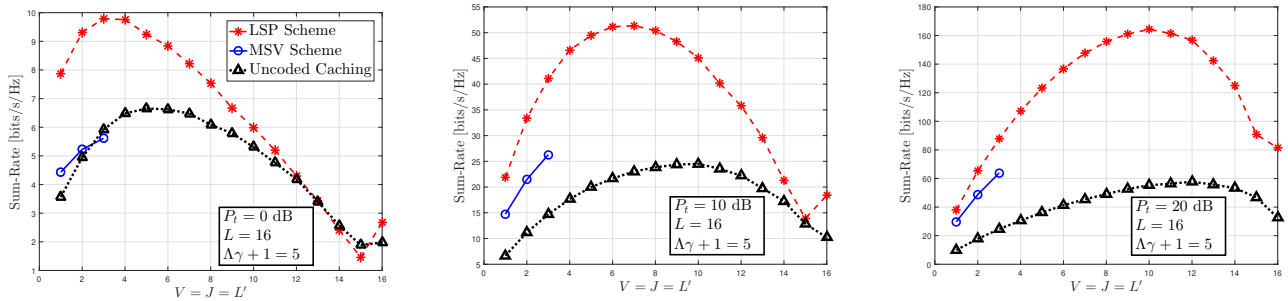

Fig. 3: Sum-rate comparisons for $P_t = 0, 10, 20$ dB, where $L = 16$ and $\Lambda\gamma + 1 = 5$

of our numerical results, while using uncoded caching as the benchmark. We use L' $(L' \leq L)$ to denote the number of simultaneously served users in uncoded caching. We present some numerical results for the sum-rate comparisons among the MSV scheme, the LSP scheme and the uncoded caching *with unicasting transmissions* in Figs. 1–3. For simplicity, we assume that the factors $\{\beta_{g,b}\}_{g\in\Lambda, b\in[B]}$, accounting for the large-scale fading, are deterministic and equal to 1, which means that the users are statistically symmetric. Moreover, the AWGN power in each received signal is normalized to 1, i.e., $N_0 = 1$. As done in [17], choosing parameter L' in uncoded caching (similarly J in LSP and V in MSV) allows the flexibility to optimize the corresponding sum-rate. This is because simultaneously serving more users will decrease the freedom to design the beamforming vector, resulting in a smaller rate to each served user. There normally exist some cutoff points between the number of served users and the single-link rate. We use the notation Δ_{LSP} (or Δ_{MSV}) to denote the sum-rate boost of the LSP scheme (or the MSV) over uncoded caching, where J (or V) and L' are separately selected for maximizing the corresponding sum-rates.

We begin the presentation of the results by considering the case $L = 12$ and $\Lambda\gamma + 1 = 4$ for different values of P_t, and where the theoretical boost, i.e. in the very high SNR, would have been $\Delta_{\text{LSP}} = (\Lambda\gamma+1) = 4 = 400\%$ and $\Delta_{\text{MSV}} = (\Lambda\gamma + L)/L = 16/12 = 133\%$. As shown in Fig. 1, by comparing the maximum sum-rates in those three cases, we can see that the LSP scheme provides a sum-rate boost Δ_{LSP} of 140% over the uncoded caching case, while the MSV scheme achieves 90% of the uncoded caching sum-rate for $P_t = 0$ dB. One interesting outcome is that unicasting outperforms multicasting (MSV) in the low-SNR regime. As P_t increases, both Δ_{LSP} and Δ_{MSV} grow. Specifically, Δ_{LSP} are 190% for $P_t = 10$ dB and 250% for $P_t = 20$ dB, respectively. In contrast, the boost provided by the MSV scheme is smaller (i.e., $\Delta_{\text{MSV}} = 110\%$ for $P_t = 10$ dB and $\Delta_{\text{MSV}} = 120\%$ for $P_t = 20$ dB).

We continue our analysis by increasing the number of antennas to 16 while retaining the same value as before for the multicasting gain. These results are depicted in Fig. 2. It is interesting to note that, while the sum-rate has increased, the rate boost of the LSP scheme (Δ_{LSP}) remains almost unchanged. On one hand, the fact that the boost remains the same comes in agreement with the DoF analysis, where this multiplicative boost is only a function of the multicasting gain and not the number of antennas. On the other hand, the fact that the multiplicative boost has remained the same shows

ISBN 978-3-8007-5686-5

that the impact of the larger antenna array is similar to both the uncoded caching delivery scheme and the LSP scheme. However, the boost of the MSV scheme becomes much worse over uncoded caching, although the corresponding sum-rate grows due to the increase of L. This implies that the increase in the number of transmit antennas is more beneficial for the uncoded caching with unicasting transmissions.

Finally, we increase the coded caching gain $\Lambda\gamma + 1$ to 5 in Fig. 3. Compared to the sum-rate boosts in Fig. 2. we can see from Fig. 3 that both of the sum-rate boosts of the LSP and the MSV increases. This is because a larger coded caching gain enables the BS to simultaneously serve more users in coded caching, thereby enhancing the sum-rate, while the increase in the coded caching gain does not impact the sum-rate of uncoded caching.

VI. Conclusion and Future Work

We have investigated two popular multi-antenna coded caching schemes, i.e., the LSP and the MSV, and designed the corresponding linear precoders for them. Specifically, the linear precoder considered here consists of two layers, where the outer-layer is for interference cancellation and the inner-layer is for beamforming. Further, we have formulated the optimization model for the inner-layer beamforming vector. In the end, we presented some numerical results that provide useful insights into the performance of these algorithms under finite SNR and finite subpacketization. The numerical comparisons have shown that the MSV scheme is not always better than uncoded caching. Specifically, uncoded caching outperforms the MSV scheme at low SNR and in large-scale antennas, while the MSV scheme is better in the large coded caching gain regime and/or in the high-SNR regime. In contrast, the LSP scheme always provides the best delivery performance than uncoded caching and the MSV scheme.

Due to the high complexity in the inner-layer beamforming vector optimization, in the future work, we will consider more simplified linear precoders (e.g., matched-filtering) and practical impacts (e.g., channel estimation errors) to analyze the delivery performance of the LSP and the MSV.

References

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

[2] 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.

[3] N. D. Sidiropoulos *et al.*, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[4] S. P. Shariatpanahi *et al.*, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[5] E. Lampiris, A. Bazco-Nogueras, and P. Elia, "Resolving the feedback bottleneck of multi-antenna coded caching," 2020, submitted to *IEEE Trans. Inf. Theory*. [Online]. Available: https://arxiv.org/abs/1811.03935

[6] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[7] E. Lampiris and P. Elia, "Achieving full multiplexing and unbounded caching gains with bounded feedback resources," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1440–1444.

[8] ——, "Adding transmitters allows unbounded coded-caching gains with bounded file sizes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1435–1439.

[9] B. Serbetci, E. Lampiris, T. Spyropoulos, and P. Elia, "Augmenting multiple-transmitter coded caching using popularity knowledge at the transmitters," in *Proc. 18th Int. Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, Jun. 2020.

[10] E. Lampiris and P. Elia, "Bridging two extremes: Multi-antenna coded caching with reduced subpacketization and CSIT," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019.

[11] ——, "Full coded caching gains for cache-less users," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7635–7651, Dec. 2020.

[12] M. Salehi *et al.*, "A multi-antenna coded caching scheme with linear subpacketization," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020.

[13] M. Salehi, A. Tölli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization-rate trade-off in multi-antenna coded caching," in *Proc. IEEE Global Commun. Conf. (Globecom)*, Dec. 2019.

[14] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Resolving the worst-user bottleneck of coded caching: Exploiting finite file sizes," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2021, pp. 1–5.

[15] ——, "Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes," 2021, submitted to *IEEE Trans. Wireless Commun.* [Online]. Available: https://arxiv.org/abs/2103.02967

[16] ——, "Coded caching gains at low SNR over Nakagami fading channels," in *Proc. 55th Asilomar Conf. Signals, Syst., Comput. (ACSSC)*, Oct./Nov. 2021, accepted for publication.

[17] E. Lampiris *et al.*, "Bridging the gap between multiplexing and diversity in finite SNR multiple antenna coded caching," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput. (ACSSC)*, Nov. 2019, pp. 1272–1277.

[18] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2091–2106, Mar. 2020.

[19] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, "Multicast beamformer design for coded caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1914–1918.

[20] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2792–2807, May 2019.

[21] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5524–5537, Oct. 2016.

[22] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design for centralized coded caching scheme," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5821–5833, Sep. 2017.

[23] M. Cheng, J. Jiang, Q. Yan, and X. Tang, "Constructions of coded caching schemes with flexible memory size," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4166–4176, Jun. 2019.

[24] L. Tang and A. Ramamoorthy, "Coded caching schemes with reduced subpacketization from linear block codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 3099–3120, Apr. 2018.

[25] P. Krishnan, "Coded caching via line graphs of bipartite graphs," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[26] H. H. S. Chittoor and P. Krishnan, "Low subpacketization coded caching via projective geometry for broadcast and D2D networks," 2019. [Online]. Available: https://arxiv.org/abs/1902.08041

[27] C. Shangguan, Y. Zhang, and G. Ge, "Centralized coded caching schemes: A hypergraph theoretical approach," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5755–5766, Aug. 2018.

[28] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Wireless coded caching with shared caches can overcome the near-far bottleneck," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 350–355.

[29] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, Jun. 2018.

[30] E. Karipidis *et al.*, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.

[31] M. Sadeghi *et al.*, "Reducing the computational complexity of multicasting in large-scale antenna systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2963–2975, May 2017.

[32] M. Sadeghi, E. Björnson, E. G. Larsson, C. Yuen, and T. L. Marzetta, "Max-min fair transmit precoding for multi-group multicasting in massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1358–1373, Feb. 2018.