Vector Coded Caching Multiplicatively Increases the Throughput of Realistic Downlink Systems

Hui Zhao, *Graduate Student Member, IEEE,* Antonio Bazco-Nogueras, *Member, IEEE,* and Petros Elia, *Member, IEEE*

***Abstract*—The recent introduction of vector coded caching has revealed that multi-rank transmissions in the presence of receiver-side cache content can dramatically ameliorate the file-size bottleneck of coded caching and substantially boost performance in error-free wire-like channels. In this work, we employ large-matrix analysis to explore the effect of vector coded caching in realistic wireless multi-antenna downlink systems. For a given downlink MISO system already optimized to exploit both multiplexing and beamforming gains, and for a fixed set of antenna and SNR resources, our analysis answers a simple question: What is the multiplicative throughput boost obtained from introducing reasonably-sized receiver-side caches that can pre-store information content? The derived closed-form expressions capture various linear precoders, and a variety of practical considerations such as power dissemination across signals, realistic SNR values, as well as feedback costs. The schemes are very simple (we simply collapse precoding vectors into a single vector), and the recorded gains are notable. For example, for 32 transmit antennas, a received SNR of 20 dB, a coherence bandwidth of 300 kHz, a coherence period of 40 ms, and under realistic file-size and cache-size constraints, vector coded caching is here shown to offer a multiplicative throughput boost of about 310% with ZF/RZF precoding and a 430% boost in the performance of already optimized MF-based (cacheless) systems. Interestingly, vector coded caching also accelerates channel hardening to the benefit of feedback acquisition, often surpassing 540% gains over traditional hardening-constrained cacheless downlink systems.**

***Index Terms*—Coded caching, linear precoding, multi-antenna transmission, random matrix analysis, downlink systems.**

I. Introduction

CACHING is widely considered to be a valuable resource toward alleviating traffic congestion in various networks [2], [3]. A particularly powerful method for exploiting cache resources can be found in the seminal work of Maddah-Ali and Niesen [4], who introduced the coded caching framework as a means for exploiting cache-aided side information at the receivers in order to remove interference. This breakthrough was originally presented for the single-stream (single-antenna), error-free, shared-link Broadcast Channel (BC), over which a central server delivers content to K cache-aided users. In this context, the server has access to a library of N files, and each user has access to their own dedicated cache of normalized size $\gamma \triangleq \frac{M}{N} \in [0,1]$ corresponding to an individual cache-size equal to the size of $M = \gamma N$ files, and corresponding to a cumulative cache size equal to $K\gamma$ times the size of the library. After a combinatorial content-allocation in each cache during the placement phase, and after each user reveals its demanded file, the delivery phase in [4] employed a novel clique-based scheme that transmitted XORs that could serve $K\gamma+1$ users at a time. This astounding multiplicative speed-up factor of $K\gamma+1$ over single-stream cacheless systems was based on the idea that a single XOR carries the desired subfiles of $K\gamma+1$ users, and that these users can utilize their own cached side information to remove undesired subfiles from the XOR in order to recover their own subfile. Unfortunately, the clique-based structure of the so-called MN coded caching scheme in [4] requires that the size of each file grows exponentially in K (cf. [5], [6]). This in turn effectively implies — under realistic file sizes — a much reduced real speedup factor $\Lambda\gamma+1 \ll K\gamma+1$ for some maximum allowed number of cache-states[1] $\Lambda \ll K$. This problem of subpacketization-constrained (or file-size constrained) coded caching is thoroughly documented in a variety of works such as [6]–[8] as well as [9]–[12].

At the same time, it also became apparent that for coded caching to develop into an impactful ingredient in wireless systems, it would have to work in conjunction with multi-antenna arrays which are rightfully recognized as the most valuable resource in modern networks. This realization brought to the fore notable research in the area of *multi-antenna coded caching* [13], [14], which considers the same model as the aforementioned cache-aided BC, except that now the server (the base-station) is endowed with multiple transmit antennas. In recent years, several related works explored various aspects of the problem, with substantial emphasis on physical-layer considerations. One of the first such works can be found in [15] which designed physical-layer adaptations of various multi-antenna coded caching schemes. Another interesting approach can be found in [16] which presented a multi-antenna coded-caching scheme for lower SNR regimes when the placement exploits prior information on the users' locations. Furthermore, the work of [17] considered the use of transmit antennas for achieving rate scalability in the limit of large K, while the work in [18], [19] nicely considered the fusion of multi-

Manuscript received February 08, 2022; revised July 26, 2022; accepted September 29, 2022. This work is supported in part by the Regional Government of Madrid through the grant 2020-T2/TIC-20710 for Talent Attraction, and by the European Research Council under the EU Horizon 2020 research and innovation program/ERC grant agreement no. 725929 (ERC project DUALITY). Part of this work was presented at the 23rd IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2022 [1]. The associate editor coordinating the review of this article and approving it for publication was Z. Zhang. *(Corresponding author: Hui Zhao.)*

Hui Zhao and Petros Elia are with the Communication Systems Department, EURECOM, 06410 Sophia Antipolis, France (email: hui.zhao@eurecom.fr; elia@eurecom.fr).

Antonio Bazco-Nogueras is with the IMDEA Networks Institute, 28918 Madrid, Spain (email: antonio.bazco@imdea.org).

[1]The cache state defines the content stored at the cache of a certain user, such that two users sharing the same cache state must store the exact same content in their caches. Having fewer cache states implies smaller subpacketization and thus smaller required file sizes, and a bounded file size forcefully reduces Λ as well as the corresponding gain $\Lambda\gamma+1$.

antenna multicast beamforming and coded caching toward improved interference management. Interesting work can also be found in [20]–[29] and in a variety of other publications. It is the case though that for most of the above schemes, the corresponding degrees-of-freedom (DoF) impact of caching was merely additive to the multiplexing gain (denoted here by Q), in the sense that in most of the above scenarios, the DoF performance stagnated at around $Q + \Lambda\gamma$ for very modest values of $\Lambda\gamma$. In essence, *due to the severity of the file-size constraint, the impact of caching was dwarfed by the existing and available multiplexing gains*, which has been extensively demonstrated in various field trials [31].

This imbalance in the impact of caching on multi-antenna systems was reversed with the introduction in [32] of vector coded caching. This reversal is owed in part to the fact that this new approach could dramatically ameliorate the subpacketization problem previously associated to XOR-based schemes. While previous multi-antenna coded caching techniques essentially focused on using multiple antennas (L transmit antennas) to efficiently deliver the aforementioned sequence of XORs of the original MN scheme, the novel method in [32] applied a decomposition-based approach that employed a clique structure *on vectors* rather than on scalars. Vector coded caching need not entail the transmission of XORs. Building on the idea of employing Λ shared caches (Λ cache states) and linear precoding, the algorithm in [32] was able to offer unprecedented performance as well as a dramatically reduced subpacketization. To be precise, for some $Q \leq L$ representing the aforementioned multiplexing gain of choice, the algorithm in [32] reduced subpacketization from being exponential in K to being exponential in K/Q, all while being able to serve up to $Q(1 + \Lambda\gamma)$ users at a time. This implied a theoretical multiplicative boost over the DoF of multiplexing-gain systems by a factor of $1 + \Lambda\gamma$, with the new DoF of $Q(1 + \Lambda\gamma)$ far exceeding the additive impact (see DoF of $Q + \Lambda\gamma$) of previous XOR-based multi-antenna coded caching approaches. It is the case though that the work in [32] focused on the error-free, asymptotically high-SNR regime, without considering any practical aspects such as power dissemination across signals, realistic SNR values, the effects of beamforming gain, or the costs of gathering channel state information (CSI). With the exception of some preliminary works like the one in [33], we know very little about the practical performance of vector coded caching in wireless systems. While this new approach was shown to be useful in an information-theoretic (DoF) sense, the real impact that this approach has on optimized downlink systems, has remained an open question.

Any attempt to establish the real impact of vector coded caching must answer a simple question: Under a fixed set of antenna and SNR resources, what is the multiplicative throughput boost obtained from being able to add receiver-side caches to downlink systems that would have otherwise been able to enjoy an optimized exploitation of multiplexing and beamforming gains. Indeed, spatial multiplexing and beamforming in multi-antenna downlink systems, and its well-studied application in the large-antenna regime or *massive* multiple-input multiple-output (MIMO) [34]–[37], is a key technology in current and future wireless networks that significantly enhances spectral efficiency. Such enhancements have been recently proven in the aforementioned field trials [31] which demonstrate that a sizeable fraction of the promising theoretic gains brought about by spatial multiplexing approaches, can indeed be attained under practical constraints.

While very considerable research has focused on a variety of advanced precoding schemes, the work-horses of spatial-multiplexing precoding are the optimized versions of linear precoding techniques such as Zero-Forcing (ZF), Regularized ZF (RZF), and Matched Filtering (MF). These techniques maintain low complexity and an ability to provide very high spectral efficiency that often comes close to the optimal performance of the non-linear Dirty-Paper Coding, especially when the number of transmit antennas L is large [34]. Furthermore, as one would expect, the acquisition of CSI is another ingredient of crucial importance in such systems, even in the presence of Time Division Duplexing (TDD) that partially reduces the CSI overhead as the dimensionality of the problem becomes larger [38]. This same CSI overhead brings to the fore the issue of channel hardening, which arises as the number of antennas increases, and which partially alleviates the stringent CSI requirements [39].

Structure of Paper and Current Contributions: The remainder of this paper is organized as follows. We introduce the system model and the considered framework in Section II. Subsequently, in Section III, we first adapt the vector coded caching approach of [32] to realistic SNR values, while considering three different linear precoding schemes: ZF, RZF and MF. After doing so, we proceed to employ random matrix theory to analyze (in Theorem 1 for MF, Theorem 2 for ZF, and Theorem 3 for RZF) the achievable throughput of vector coded caching for the three aforementioned precoders. This analysis — which naturally incorporates the standard cacheless case corresponding to $\gamma = 0$ — captures any SNR and any number of users.

Subsequently, based on the derived asymptotic performance, in Section IV we optimize both the cacheless as well as the cache-aided algorithms by accounting for the CSI acquisition costs, and by optimizing over the total number of simultaneously served streams (users). This optimization, which is performed as a function of SNR, of L and of the CSI acquisition costs, can be found in Theorems 4, 5. The same optimization yields systems that are separately calibrated to better balance multiplexing gains with beamforming gains, in the presence or absence of caching. In this same section we also derive the ratio between the throughputs of the (independently) optimized cache-aided and cacheless systems. This ratio represents the multiplicative throughput boost offered by caching, over optimized cacheless downlink systems with the same power and antenna resources. Subsequently, in Section V we numerically verify the accuracy of the derived expressions, showing that they characterize very precisely the actual performance. This evaluation allows us to demonstrate the substantial gains from using caching, highlighting realistic regimes of SNR, L, CSI costs, file sizes and cache sizes. In Section VI we present the main conclusions, while in the appendices we host some of the remaining proofs.

Notations: $\mathbb{C}$ stands for the set of complex numbers,

$\mathbf{I}_L \in \mathbb{C}^{L\times L}$ denotes the $L\times L$ identity matrix, and $\mathbf{0}_L \in \mathbb{C}^{L\times 1}$ denotes the all-zero vector. We use $X \sim \mathcal{Y}$ to denote that X follows the statistical distribution $\mathcal{Y}$. Furthermore, $|\cdot|$ denotes either the cardinality of a set or the magnitude of a complex number, $||\cdot||$ denotes the norm-2 operator for a vector, while we also define $[Z] \triangleq \{1, 2, \cdots, Z\}$ for a positive integer Z. Additionally, $\mathrm{Tr}\{\cdot\}$ and $\mathbb{E}\{\cdot\}$ denote the trace and the expectation operators, respectively, whereas $(\cdot)^T$, $(\cdot)^*$ and $(\cdot)^H$ denote the non-conjugate transpose, conjugate part and conjugate transpose of a matrix, respectively. In asymptotic analysis, $f(x) = o(g(x))$ as $x \to \infty$ denotes that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$. $\xrightarrow{a.s.}$ stands for almost sure convergence. If $X \xrightarrow{a.s.} \mathring{X}$ and $\mathring{X}$ is deterministic, we call $\mathring{X}$ the asymptotic deterministic equivalent of X. Moreover, in the limit of $x \to \infty$, our use of $A(x) \doteq B(x)$ will mean that $A(x) = B(x) + o(1)$.

II. System Model and Problem Description

A. System Model

We consider a downlink MISO scenario where an L-antenna base station (BS) serves K single-antenna cache-aided users. The BS has access to a library of N equally-sized files, and each user is endowed with a local memory (or cache) of size equal to the size of M library files ($M < N$), such that each user can store a fraction $\gamma = \frac{M}{N} \in [0, 1)$ of the library content. We denote the library content by $\mathcal{F}$ and the n-th file by W_n, such that $\mathcal{F} \triangleq \{W_n\}_{n=1}^N$.

We consider the wireless channel to be modeled as a symmetric Rayleigh fading channel, where all channel coefficients are assumed to be independent and identically distributed (i.i.d.). When describing a general transmission, our notation will often incorporate the subset $\mathcal{K} \subseteq [K]$ of users that are simultaneously served during that transmission. Consequently, in our communication model, the received signal at the k-th user in $\mathcal{K}$ is given by

$$y_{\mathcal{K}(k)} = \mathbf{h}_{\mathcal{K}(k)}^T \mathbf{x}_{\mathcal{K}} + z_{\mathcal{K}(k)}, \tag{1}$$

where $k \in [|\mathcal{K}|]$, where $z_{\mathcal{K}(k)} \in \mathbb{C}$ represents the corresponding Additive White Gaussian Noise (AWGN) with zero-mean and unit-variance, where $\mathbf{x}_{\mathcal{K}} \in \mathbb{C}^{L\times 1}$ denotes the transmitted signal vector that simultaneously serves the users in $\mathcal{K}$, and where $\mathbf{h}_{\mathcal{K}(k)} \in \mathbb{C}^{L\times 1}$ represents the channel vector for the channel from the BS to the k-th user in $\mathcal{K}$. As mentioned, $\mathbf{h}_{\mathcal{K}(k)}$ is assumed to be an i.i.d. Gaussian random vector with mean $\mathbf{0}_L$ and covariance matrix $\mathbf{I}_L$. Finally, $\mathbf{x}_{\mathcal{K}}$ is obtained by applying a specific precoding scheme (which we will detail later on) to the information vector $\mathbf{s}_{\mathcal{K}} \in \mathbb{C}^{|\mathcal{K}|\times 1}$ intended for the users in $\mathcal{K}$, where $\mathbf{s}_{\mathcal{K}}$ has mean $\mathbf{0}_{|\mathcal{K}|}$ and covariance matrix $\mathbf{I}_{|\mathcal{K}|}$.

We consider an average power normalization, where the power is averaged over both transmit symbols and channel realizations, i.e., $\mathbb{E}\{||\mathbf{x}_{\mathcal{K}}||^2\} \leq P_t$, where P_t is the average power constraint. As is common in practical downlink settings, we assume TDD uplink-downlink transmissions, such that the BS estimates the downlink channels through uplink pilot transmissions by applying channel reciprocity.

We proceed to describe the main structure of the scheme, first doing so without specifying the linear precoding class that is used. We will also formally define the main performance metrics investigated in this paper.

B. Signal-Level Vector Coded Caching for Finite SNR

Building on the general vector-clique structure in [32], we are here free to choose the precoding schemes, as well as calibrate at will the dimensionality of each vector clique. This freedom is essential in controlling the impact of CSI costs and of power-splitting across users, both of which directly affect the performance in practical SNR regimes.

We proceed to describe the cache placement phase and the subsequent delivery phase.

1) Placement Phase: The first step involves the partition of each library file W_n into $\binom{\Lambda}{\Lambda\gamma}$ non-overlapping equally-sized subfiles $\left\{W_n^{\mathcal{T}} : \mathcal{T} \subseteq [\Lambda], |\mathcal{T}| = \Lambda\gamma\right\}$, each labeled by some $\Lambda\gamma$-tuple $\mathcal{T} \subseteq [\Lambda]$. As discussed in Section I, the number of cache states Λ is chosen to satisfy the file-size constraint; in our case, the subpacketization is $\binom{\Lambda}{\Lambda\gamma}$, which naturally serves as a lower bound on the file sizes. Subsequently the K users are *arbitrarily* separated into Λ disjoint groups $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_\Lambda$, where the g-th group, which consists of $B = \frac{K}{\Lambda}$ users[2], is given by $\mathcal{D}_g \triangleq \left\{b\Lambda + g\right\}_{b=0}^{B-1} \subseteq [K]$. The ϑ-th user of this g-th group is denoted by $\mathrm{U}_{g,\vartheta}$.

At this point, all the users belonging to the same group are assigned the same cache state and thus proceed to cache *identical* content. In particular, for those in the g-th group, this content takes the form $\mathcal{Z}_{\mathcal{G}_g} = \left\{W_n^{\mathcal{T}} : \mathcal{T} \ni g, \forall n \in [N]\right\}$. This grouping as well as the entire placement phase, are naturally done before the users' requests take place, and of course well before the channel states are known to the BS.

2) Delivery Phase: This phase starts when each user $\kappa \in [K]$ simultaneously asks for its intended file, denoted here by W_{d_κ}, $d_\kappa \in [N]$. The BS selects Q users from each group, where $Q \leq B$ is a variable that will be optimized afterwards and which is the equivalent of the multiplexing gain. By doing so, the BS decides to first 'encode' over the first ΛQ users, and to repeat the encoding process B/Q times[3]. To deliver to the ΛQ users, the transmitter employs $\binom{\Lambda}{\Lambda\gamma+1}$ sequential transmission stages. During each such stage, the BS simultaneously serves a unique set Ψ of $|\Psi| = \Lambda\gamma + 1$ groups, corresponding to a total of $Q(\Lambda\gamma + 1)$ users served at a time (i.e., per stage). At the end of the $\binom{\Lambda}{\Lambda\gamma+1}$ transmission stages, all the ΛQ users obtain their intended files. By repeating this process $\lceil \frac{B}{Q} \rceil$ times, all the K users obtain their intended files. As suggested above, the factor $G \triangleq \Lambda\gamma + 1$ describes the number of user groups that are simultaneously served. Another crucial parameter includes the multiplexing gain Q which, unlike in [32], will be here subject to optimization.

For example, let us consider a setting with $G = 3$, $B = 4$, and $\Lambda = 40$, and a choice of $Q = 2$. The delivery will be split into $\frac{B}{Q} = 2$ encoding processes, and each process is split into

[2]For clarity of exposition, and without limiting the scope of the results, we will consider K to be a multiple of Λ. The general case can be readily handled (cf. [32]), and in Section V we provide a related example.

[3]To clarify, what the above says is the following. If there are, e.g., $B = 2Q$ users per group and thus $K = 2\Lambda Q$ users in total, then the algorithm that we describe here will be first applied to the first ΛQ users, and then, after this delivery is done, the same algorithm will apply to the remaining ΛQ users, thus eventually satisfying all K users. Also note that a small amount of additional subpacketization can easily resolve the case where B/Q may not be an integer.

$\binom{\Lambda}{G}$ so-called stages. Each stage will involve the transmission to a different set (triplet in this case) of cache groups $\Psi' \subseteq [\Lambda]$ where $|\Psi'| = G = 3$. In each such stage, the BS serves $Q = 2$ users from each of the above three cache groups, which allows for serving $GQ = 6$ users at a time. For example, the first stage can correspond to the set $\Psi = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$. The difference between the two processes is that in the first delivery process, the BS serves the first $Q = 2$ users in each cache-group, while in the second process, the BS serves the last $Q = 2$ users in each cache-group (cf. Fig. 1). We refer to the users currently served in a delivery process as active users and to all the other users as passive users in Fig. 1.

Let us now focus on a single transmission stage. As mentioned above, at each such stage, we pick a set $\Psi \subseteq \Lambda$ of $G = \Lambda\gamma + 1$ groups that will be served simultaneously. From within these chosen groups, we will serve $Q \leq B$ users per group. In particular, for each user $\mathrm{U}_{\psi,\vartheta}$ of some group $\psi \in \Psi$, this stage will deliver all subfiles[4] $s_{\psi,\vartheta}$ by transmitting

$$\mathbf{x}_\Psi = \frac{1}{\sqrt{G}} \sum\nolimits_{\psi\in\Psi} \rho_\psi \sum\nolimits_{\vartheta=1}^{Q} \mathbf{v}_{\psi,\vartheta} s_{\psi,\vartheta}, \tag{2}$$

where $\mathbf{v}_{\psi,\vartheta} \in \mathbb{C}^{L\times 1}$ denotes the precoder applied to the subfile intended by user $\mathrm{U}_{\psi,\vartheta}$, and where ρ_ψ denotes the power normalization factor for group $\psi \in \Psi$, applied under a total power constraint P_t. Upon defining $\mathbf{V}_\psi \in \mathbb{C}^{L\times Q}$ as $\mathbf{V}_\psi \triangleq [\mathbf{v}_{\psi,1}| \dots |\mathbf{v}_{\psi,Q}]$ and $\mathbf{s}_\psi \in \mathbb{C}^Q$ as $\mathbf{s}_\psi \triangleq [s_{\psi,1}, \dots, s_{\psi,Q}]^T$, the above takes the simple form

$$\mathbf{x}_\Psi = \frac{1}{\sqrt{G}} \sum\nolimits_{\psi\in\Psi} \rho_\psi \mathbf{V}_\psi \mathbf{s}_\psi. \tag{3}$$

Remark 1. *It is easy to see that the described scheme simply involves a carefully selected linear combination of G linear-precoding vectors that are now to be sent simultaneously. It is also easy to see that the above scheme also incorporates the traditional cacheless downlink scenario corresponding to $\gamma = 0$, which itself corresponds to $G = |\Psi| = 1$. In such case, the transmit signal expression reverts to the simpler common expression $\mathbf{x} = \rho\mathbf{V}\mathbf{s}$.*

For decoding to work, the subfiles must be chosen carefully. This choice follows the principles of coded caching, and in particular of vector coded caching. Thus, when considering the transmission stage which serves the $G = \Lambda\gamma + 1$ groups in Ψ, the subfile transmitted to user $\mathrm{U}_{\psi,\vartheta}$ is here selected to be $W_{d_{\psi,\vartheta}}^{\Psi\setminus\{\psi\}}$, simply because this subfile is stored in the cache of each user of every other group in Ψ except ψ. Because of this structure, the users of a particular group can remove the inter-group interference from the other $\Lambda\gamma$ groups by using their cached content. On the other hand, following the principles of vector coded caching, the intra-group interference is handled with linear precoding that 'separates' the signals of the users from the same group. Naturally one can imagine that cache-aided removal of interference as well as 'nulling out' of interference, both require knowledge of the composite precoder-channel coefficients (cf. (4) and (5)). These so-called *composite CSI* costs will be explicitly accounted for in our analysis. We proceed to elaborate on the precoders and the transmissions.

[4]In a slight abuse of notation, we use the term "subfile" to refer both to the actual subfile generated after file-splitting, as well as to the corresponding complex-valued information symbol $s_{\psi,\vartheta}$.

C. Vector Coded Caching for the Physical Layer

We now emphasize on the physical layer details of the communication scheme. Our description will focus on the transmission that serves a specific set Ψ of user-groups. First let us recall that $\mathbf{V}_\psi \in \mathbb{C}^{L\times Q}$ denotes the precoding matrix for the symbols of users in group $\psi \in \Psi$. We note that, as is common, our analysis will assume Gaussian signaling. Then let us note that for an average power constraint P_t, the power normalization factor ρ_ψ from (3), takes the form $\rho_\psi = \sqrt{\frac{P_t}{\mathbb{E}\{\mathbf{s}_\psi^H \mathbf{V}_\psi^H \mathbf{V}_\psi \mathbf{s}_\psi\}}} = \sqrt{\frac{P_t}{\mathbb{E}\{\mathrm{Tr}\{\mathbf{V}_\psi^H \mathbf{V}_\psi\}\}}}$. Then the subsequent corresponding received signal at user $\mathrm{U}_{\psi,k}$ (i.e., at the k-th user of group $\psi \in \Psi$), will take the form

$$y_{\psi,k} = \frac{\mathbf{h}_{\psi,k}^T}{\sqrt{G}} \rho_\psi \mathbf{V}_\psi \mathbf{s}_\psi + \underbrace{\frac{\mathbf{h}_{\psi,k}^T}{\sqrt{G}} \sum\nolimits_{\substack{\phi\in\Psi \\ \phi\neq\psi}} \rho_\phi \mathbf{V}_\phi \mathbf{s}_\phi}_{\text{inter-group interference}} + z_{\psi,k}. \tag{4}$$

As previously mentioned, the inter-group interference[5] experienced by user $\mathrm{U}_{\psi,k}$, can be removed from $y_{\psi,k}$ by exploiting that same user's cached content and that user's composite CSI $\{\mathbf{h}_{\psi,k}^T \mathbf{v}_{\phi,k'} \rho_\phi\}_{\phi\in\{\Psi\setminus\psi\},\, k'\in[Q]}$. Then, after the cache-aided removal of this inter-group interference, the equivalent received signal at $\mathrm{U}_{\psi,k}$ is given by

$$y'_{\psi,k} = \frac{\rho_\psi}{\sqrt{G}} \mathbf{h}_{\psi,k}^T \mathbf{v}_{\psi,k} s_{\psi,k} + \underbrace{\frac{\rho_\psi}{\sqrt{G}} \sum\nolimits_{\vartheta=1,\vartheta\neq k}^{Q} \mathbf{h}_{\psi,k}^T \mathbf{v}_{\psi,\vartheta} s_{\psi,\vartheta}}_{\text{intra-group interference}} + z_{\psi,k}. \tag{5}$$

Consequently, the corresponding SINR for information decoding at $\mathrm{U}_{\psi,k}$, is given by

$$\mathrm{SINR}_{\psi,k} = \frac{\frac{\rho_\psi^2}{G} |\mathbf{h}_{\psi,k}^T \mathbf{v}_{\psi,k}|^2}{1 + \frac{\rho_\psi^2}{G} \sum_{\vartheta=1,\vartheta\neq k}^{Q} |\mathbf{h}_{\psi,k}^T \mathbf{v}_{\psi,\vartheta}|^2}. \tag{6}$$

On the other hand, in the cacheless case of $\gamma = 0$, the received signal $y_k = \rho \mathbf{h}_k^T \mathbf{v}_k s_k + \rho \sum_{\vartheta=1,\vartheta\neq k}^{Q} \mathbf{h}_k^T \mathbf{v}_\vartheta s_\vartheta + z_k$ at some user k naturally carries no inter-group interference (as there are no other groups to simultaneously serve), and the SINR takes the standard form $\mathrm{SINR}_k = \frac{\rho^2 |\mathbf{h}_k^T \mathbf{v}_k|^2}{1+\rho^2 \sum_{\vartheta=1,\vartheta\neq k}^{Q} |\mathbf{h}_k^T \mathbf{v}_\vartheta|^2}$. Therefore, the instantaneous rate

$$R_{\psi,k} = \ln\left(1 + \mathrm{SINR}_{\psi,k}\right) \qquad \text{nats/s/Hz} \tag{7}$$

for user $\mathrm{U}_{\psi,k}$ is obtained by evaluating the above, at the SINR value in (6).

We consider the MF, ZF and RZF linear precoding schemes, selected here for being very common, simple, as well as competitive in terms of rate performance [41], [42]. As is

[5]As a reminder, the term inter-group interference refers to the received signal component whose power is due to the information meant for users originating from other groups.

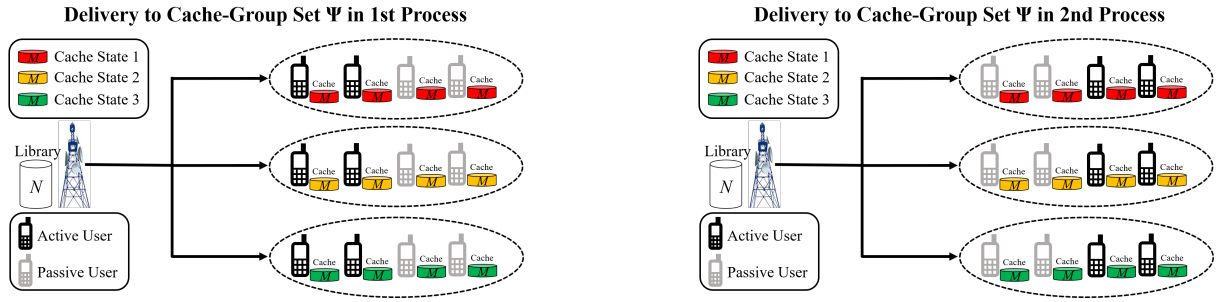


Fig. 1: An example of vector coded caching with $G = 3$, $\Lambda = 40$, $B = 4$ and $Q = 2$.

known, the corresponding precoding matrices $\mathbf{V}_\psi$ take the form:

$$\mathbf{V}_\psi = \begin{cases} \mathbf{H}_\psi^H, & \text{MF Precoder} \\ \mathbf{H}_\psi^H \left(\mathbf{H}_\psi \mathbf{H}_\psi^H\right)^{-1}, & \text{ZF Precoder} \\ \mathbf{H}_\psi^H \left(\mathbf{H}_\psi \mathbf{H}_\psi^H + \alpha \mathbf{I}_Q\right)^{-1}, & \text{RZF Precoder,} \end{cases} \tag{8}$$

where $\mathbf{H}_\psi \triangleq \left[\mathbf{h}_{\psi,1}|\mathbf{h}_{\psi,2}|\cdots|\mathbf{h}_{\psi,Q}\right]^T \in \mathbb{C}^{Q\times L}$ denotes the channel matrix for the channel from the BS to the Q chosen users belonging to group $\psi \in \Psi$, and where α is the regularization factor of the RZF precoder [41]. It is worth recalling that the RZF precoder reverts to the ZF precoder when $\alpha = 0$, and to the MF precoder when $\alpha \to \infty$, and also that Q is bounded above by B and, in the case of the ZF/RZF precoding, it is also bounded as $Q \leq L$. For simplicity we assume that $\alpha = L/P_t$, which is a commonly used assumption throughout the literature [41], [43], [44].

We will henceforth use the term (G, Q)-vector coded caching, to refer to the vector coded caching scheme when it serves G groups with Q users per group. We will also use the term *MF-based (G, Q)-vector coded caching* to refer to the same scheme when the underlying precoder is MF, and similarly we will use ZF-based or RZF-based (G, Q)-vector coded caching, for the other two precoders. Let us now formally define some important metrics of interest.

Definition 1. (Average sum-rate and effective sum-rate). *For a (G, Q)-vector coded caching scheme, its average sum-rate is denoted by $\bar{R}(G, Q)$ and is defined as the total data-transmission rate (before accounting for CSI costs) summed over the GQ simultaneously served users, and averaged over the fading. Similarly, the* effective average sum-rate $\bar{\mathcal{R}}(G, Q)$ *will represent the corresponding average rate after through all CSI costs are duly accounted for.*

Definition 2. (Effective gain over MISO). *For a given set of L and SNR resources, and a fixed underlying precoder class, the effective gain, after accounting for CSI costs, of the (G, Q)-vector coded caching over the cacheless scenario (corresponding to $G = 1$, and an operating multiplexing gain Q') will be denoted as $\mathcal{G}(G, Q; 1, Q') \triangleq \frac{\bar{\mathcal{R}}(G,Q)}{\bar{\mathcal{R}}(1,Q')}$ in the form of the ratios of the* effective rates.

III. Analysis of the Average Rate and of the Effective Gain over MISO

In this section, we analyze the average sum-rates and the corresponding effective rates achieved by the cache-aided downlink schemes of Section II-B for the MF, ZF and RZF linear precoders of interest. After doing so, we also report the effective gains offered by these (G, Q)-vector coded caching schemes, over the $(G = 1, Q')$ cacheless equivalents.

We will henceforth consider the ratio $c \triangleq Q/L$, while we will often use the notation $c' \triangleq Q'/L$ when referring explicitly to the cacheless equivalent. The two ratios can be chosen independently. When applying large matrix analysis, we will be assuming a fixed $c > 0$ and a fixed $c' > 0$.

A. MF Precoding

To derive the average sum-rate of vector coded caching with MF precoding, we first recall that the elements of $\mathbf{H}_\psi$ are i.i.d. Gaussian random variables with zero mean and unit variance, which implies that $\mathbb{E}\left\{\mathrm{Tr}\left\{\mathbf{H}_\psi \mathbf{H}_\psi^H\right\}\right\} = LQ$ (cf. [45]), which then implies that the power normalization factor ρ_ψ takes the form $\rho_\psi = \sqrt{P_t/\mathbb{E}\{\mathrm{Tr}\{\mathbf{H}_\psi \mathbf{H}_\psi^H\}\}} = \sqrt{\frac{P_t}{QL}}$ (cf. [46]). This in turn yields (cf. (8), (3)) a transmitted signal of the form

$$\mathbf{x}_\Psi = \sqrt{\tfrac{P_t}{GQL}} \sum_{\psi\in\Psi} \mathbf{H}_\psi^H \mathbf{s}_\psi = \sqrt{\tfrac{P_t}{GQL}} \sum_{\psi\in\Psi} \sum_{\vartheta=1}^{Q} \mathbf{h}_{\psi,\vartheta}^* s_{\psi,\vartheta}. \tag{9}$$

The corresponding average sum-rate is presented below. We note that Theorem 1 focuses on the case of $Q > 1$. However, the analysis for $Q = 1$ is straightforward and follows the same large-matrix properties and principles. The only difference is that for the single-stream scenario, one can deviate from the current scheme, and employ XORs rather than linear combinations over the complex numbers. This is not covered in our work here.

Theorem 1. *For any given P_t and $c = Q/L$, the average sum-rate $\bar{R}^{\mathrm{MF}}$ of the MF-based (G, cL)-vector coded caching scheme in the large L regime satisfies*

$$\bar{R}^{\mathrm{MF}}(G, cL) \doteq c\, GL \ln\left(1 + \frac{1}{c}\frac{P_t}{P_t + G}\right). \tag{10}$$

Proof. The proof can be found in Appendix I. □

The following directly distils the above result to the cacheless case.[6]

Corollary 1. *In the limit of large L, and for any fixed P_t and c', the average sum-rate of the (traditional, cacheless) MF-based MISO BC with $c'L$ streams satisfies*

$$\bar{R}^{\mathrm{MF}}(1, c'L) \doteq c'\ L \ln\left(1 + \frac{1}{c'}\frac{P_t}{P_t+1}\right). \tag{11}$$

B. ZF Precoding

Moving now to the case of ZF-based vector coded caching, and focusing again on a set of groups Ψ and on the transmission stage corresponding to some group $\psi \in \Psi$, the power control factor takes the form $\rho_\psi^2 = \frac{P_t}{\mathbb{E}\{\mathrm{Tr}\{(\mathbf{H}_\psi \mathbf{H}_\psi^H)^{-1}\}\}}$, while the transmitted signal from (3) becomes

$$\mathbf{x}_\Psi = \frac{1}{\sqrt{G}} \sum_{\psi\in\Psi} \rho_\psi \mathbf{H}_\psi^H \left(\mathbf{H}_\psi \mathbf{H}_\psi^H\right)^{-1} \mathbf{s}_\psi. \tag{12}$$

This in turn yields a received signal at user $\mathrm{U}_{\psi,k}$ which — after the cache-aided removal of the inter-group interference (cf. (4)) — takes the form

$$\begin{aligned} y'_{\psi,k} &= \frac{1}{\sqrt{G}} \rho_\psi \mathbf{h}_{\psi,k}^T \mathbf{H}_\psi^H \left(\mathbf{H}_\psi \mathbf{H}_\psi^H\right)^{-1} \mathbf{s}_\psi + z_{\psi,k} \\ &= \frac{1}{\sqrt{G}} \rho_\psi \left(\mathbf{1}_k^T \mathbf{H}_\psi\right) \mathbf{H}_\psi^H \left(\mathbf{H}_\psi \mathbf{H}_\psi^H\right)^{-1} \mathbf{s}_\psi + z_{\psi,k} \\ &= \frac{1}{\sqrt{G}} \rho_\psi s_{\psi,k} + z_{\psi,k}, \end{aligned} \tag{13}$$

where $\mathbf{1}_k \in \mathbb{C}^{Q\times 1}$ denotes the vector whose components are all zero except for the k-th element, which equals 1. After considering that all intra-group interference is canceled by means of ZF precoding, the SINR at user $\mathrm{U}_{\psi,k}$ is given by

$$\mathrm{SINR}_{\psi,k}^{\mathrm{ZF}} = \frac{P_t}{G\,\mathbb{E}\left\{\mathrm{Tr}\left\{\left(\mathbf{H}_\psi \mathbf{H}_\psi^H\right)^{-1}\right\}\right\}}. \tag{14}$$

With this in place, we proceed with the following theorem.

Theorem 2. *For $c = \frac{Q}{L} \in (0,1)$, the average sum-rate $\bar{R}_{\mathrm{sum}}^{\mathrm{ZF}}$ of the ZF-based (G,Q)-vector coded caching scheme takes the form*

$$\bar{R}^{\mathrm{ZF}}(G,Q) = QG \ln\left(1 + \frac{P_t}{G}\left(\frac{1}{c} - 1\right)\right). \tag{15}$$

Proof. Directly from [50], and from the fact that $\mathbf{H}_\psi \mathbf{H}_\psi^H$ is a Wishart matrix with L degrees of freedom, we know that $\mathbb{E}\left\{\mathrm{Tr}\left\{\left(\mathbf{H}_\psi \mathbf{H}_\psi^H\right)^{-1}\right\}\right\} = \frac{Q}{L-Q}$ for $L > Q$. Naturally, $\mathrm{SINR}_{\psi,k}^{\mathrm{ZF}}$ is deterministic and constant across all simultaneously served users. By summing the average rate of each of the GQ served users, we obtain (15). □

[6]It is worth noting that while there have been various works (cf. [44], [46]–[48]) analyzing the MF sum-rate in traditional massive MIMO systems, the result derived in this work here entails less assumptions. For example, focusing on the large-L regime, the result in [47] directly assumes a tight Jensen's bound, while the result in [48] is under a so-called "near deterministic" assumption in low/high SNRs. On the other hand, our method here draws from the uplink analysis in [49], and only employs a large-L assumption to derive the exact asymptotic optimality for any value of SNR.

C. RZF Precoding

We finally consider our third precoder, and do so in the asymptotic regime of large L and fixed c. We first note that the received signal at $\mathrm{U}_{\psi,k}$ — after cache-aided removal of the inter-group interference — takes the form

$$y'_{\psi,k} = \frac{\rho_\psi}{\sqrt{G}} \sum_{\vartheta=1}^{Q} \mathbf{h}_{\psi,k}^T \left(\alpha \mathbf{I}_L + \mathbf{H}_\psi^H \mathbf{H}_\psi\right)^{-1} \mathbf{h}_{\psi,\vartheta}^* s_{\psi,\vartheta} + z_{\psi,k}. \tag{16}$$

For $\mathbf{H}_{\psi,-k}$ denoting the matrix resulting from $\mathbf{H}_\psi$ after removing its k-th row, we can define

$$A_{\psi,k} \triangleq \mathbf{h}_{\psi,k}^T \Big(\alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k}\Big)^{-1} \mathbf{h}_{\psi,k}^*, \tag{17}$$

$$\begin{aligned} B_{\psi,k} \triangleq\ & \mathbf{h}_{\psi,k}^T \left(\alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k}\right)^{-1} \mathbf{H}_{\psi,-k}^H \\ & \times \mathbf{H}_{\psi,-k} \big(\alpha \mathbf{I}_L + \mathbf{H}_{\psi,-k}^H \mathbf{H}_{\psi,-k}\big)^{-1} \mathbf{h}_{\psi,k}^*. \end{aligned} \tag{18}$$

With these notations, we can derive the SINR at user $\mathrm{U}_{\psi,k}$ as

$$\mathrm{SINR}_{\psi,k}^{\mathrm{RZF}} = \frac{A_{\psi,k}^2 \frac{\rho_\psi^2}{G}}{\big(1 + A_{\psi,k}\big)^2 + \frac{\rho_\psi^2}{G} B_{\psi,k}}, \tag{19}$$

where the proof of (19) is relegated to Appendix II-A.

We can now present the asymptotic deterministic equivalent of the sum-rate of our proposed scheme when RZF is applied. We recall that in the limit of large L, the deterministic value $\mathring{X}$ represents the *asymptotic deterministic equivalent* of X if $X \xrightarrow{a.s.} \mathring{X}$.

Theorem 3. *In the large-L regime with fixed $c = Q/L$, the average sum-rate $\bar{R}^{\mathrm{RZF}}$ of RZF-based (G,Q)-vector coded caching takes the form*

$$\begin{aligned} \bar{R}^{\mathrm{RZF}}(G,Q) &\doteq \mathring{R}^{\mathrm{RZF}}(G, cL) \\ &\triangleq c\,GL \ln\left(1 + \frac{a_{\psi,k}^2 p_\psi^2 / G}{\big(1+a_{\psi,k}\big)^2 + P_t/G}\right), \end{aligned} \tag{20}$$

where $\mathring{R}^{\mathrm{RZF}}$ is the deterministic equivalent[7] of $\bar{R}^{\mathrm{RZF}}$ and

$$\begin{aligned} a_{\psi,k} &\triangleq \tfrac{1}{2}\left[\sqrt{(1-c)^2 P_t^2 + 2(1+c)P_t + 1} + (1-c)P_t - 1\right] \\ p_\psi^2 &\triangleq \frac{P_t}{a_{\psi,k} - \frac{P_t}{2}\left(\frac{P_t(c-1)^2 + c + 1}{\sqrt{P_t^2(c-1)^2 + 2(c+1)P_t + 1}} + 1 - c\right)}. \end{aligned} \tag{21}$$

Proof. The proof is based on the derivation of the asymptotic deterministic equivalent of the SINR, and it is presented in Appendix II. □

D. Accounting for the CSI Costs

To account for the cost of CSI acquisition under TDD, we consider a basic CSI-acquisition effort where at the beginning of each transmission stage, the GQ served users send uplink orthogonal pilot symbols, from which the BS can estimate the downlink channel matrix, under the assumption of channel reciprocity. Then the CSI-acquisition process engages downlink training, of similar complexity, in order to communicate the

[7]This entails a small abuse of terminology, as it is $\mathring{R}^{\mathrm{RZF}}/L$ that is the deterministic equivalent of $\bar{R}^{\mathrm{RZF}}/L$.

composite CSI that here allows our receivers to perform cache-aided cancellation of the inter-group interference (cf. (4)) from their signal. This acquisition process for gathering composite CSI, with the same aforementioned complexity per served user, is standard in a variety of traditional communications techniques such as SIC-based approaches. For additional details, please refer to [51]. To account for this CSI-acquisition overhead, we directly extend the commonly-used approach in [52]–[55], that easily allows us to calculate the effective average sum-rate (cf. Definition 1) for each precoder $i \in \{\text{MF, ZF, RZF}\}$, to be

$$\bar{\mathcal{R}}^i = \left(1 - \frac{\beta_{\text{tot}} GQ}{T_c W_c}\right) \bar{R}^i = (1 - c\ \zeta_{G,Q})\, \bar{R}^i, \qquad (22)$$

where β_{tot} is the number of resources per user and per block used for pilot transmission, $\bar{R}^i$ is the previously calculated average sum-rate before accounting for CSI costs, where T_c and W_c are the coherence time and coherence bandwidth, respectively, and where $\zeta_{G,Q} \triangleq \frac{\beta_{\text{tot}} GL}{T_c W_c}$. For completeness we report the effective rates in the following corollary. The proof is direct as it merely involves applying (22) in the expressions from Theorems 1–3. We recall that $a_{\psi,k}$ and p_ψ are defined in Theorem 3.

Corollary 2. *The effective rates of the proposed vector coded caching schemes under MF, ZF, and RZF precoding take the form, respectively,*

$$\bar{\mathcal{R}}^{MF}(G,Q) \doteq (1 - c\ \zeta_{G,Q})\, c\ GL \ln\left(1 + \frac{1}{c}\frac{P_t}{P_t + G}\right)$$

$$\bar{\mathcal{R}}^{ZF}(G,Q) = (1 - c\ \zeta_{G,Q})\, QG \ln\left(1 + \frac{P_t}{G}\left(\frac{1}{c} - 1\right)\right)$$

$$\bar{\mathcal{R}}^{RZF}(G,Q) \doteq (1 - c\ \zeta_{G,Q}) c GL \ln\left(1 + \frac{a_{\psi,k}^2 p_\psi^2 / G}{(1 + a_{\psi,k})^2 + {}^{P_t}/_{G}}\right)$$

E. Effective Gains over Cacheless MISO Systems

At this point, with Theorems 1, 2, 3 in place, and in conjunction with Corollary 2, we can directly report the effective gains over cacheless MISO. For each of the three precoder classes, MF, ZF, and RZF, and for a fixed set of antenna and SNR resources, we obtain the effective gain $\mathcal{G}(G,Q;1,Q') = \frac{\bar{\mathcal{R}}(G,Q)}{\bar{\mathcal{R}}(1,Q')}$ (cf. Definition 2) of the (G,Q)-vector coded caching schemes over the cacheless scenario ($G = 1$) with some chosen number of streams Q'. These effective gains are collected together in the following corollary.

Corollary 3. *The effective gains of the proposed vector coded caching schemes under MF, ZF, and RZF precoding satisfy, respectively,*

$$\mathcal{G}_{\text{MF}}(G,Q;1,Q') \doteq \xi \frac{GQ}{Q'} \frac{\ln\left(1 + \frac{L}{Q}\frac{P_t}{P_t+G}\right)}{\ln\left(1 + \frac{L}{Q'}\frac{P_t}{P_t+1}\right)}$$

$$\mathcal{G}_{\text{ZF}}(G,Q;1,Q') = \xi \frac{GQ}{Q'} \frac{\ln\left(1 + \frac{P_t}{G}\left(\frac{L}{Q} - 1\right)\right)}{\ln\left(1 + P_t\left(\frac{L}{Q'} - 1\right)\right)}$$

$$\mathcal{G}_{\text{RZF}}(G,Q;1,Q') \xrightarrow{a.s.} \xi \frac{\mathring{R}^{\text{RZF}}(G, cL)}{\mathring{R}^{\text{RZF}}(1, c'L)}$$

where $\xi \triangleq \frac{(L - Q\zeta_{G,Q})}{(L - Q'\zeta_{1,Q'})}$ *and* $\mathring{R}^{\text{RZF}}(\cdot,\cdot)$ *is defined in* (20).

IV. Optimizing Physical Layer Vector Coded Caching

Theorems 1–3 reveal the important dependence of vector coded caching on the number of streams, Q, that we choose to activate. This dependence strikes at the very core of the problems stemming from power-splitting and CSI overheads. Indeed, while an increased $Q \leq L$ allows for a higher DoF at lower subpacketization, this increase in the number of streams may not be beneficial in practice as it entails less power per stream as well as more CSI to be communicated.

For this reason, we here proceed to analytically optimize our schemes over the choices of Q. This optimization is tractable partly due to the simplicity of the achievable-rate expressions derived in the previous theorems[8], and while some of these expressions involve asymptotic approximations, they will, as we will verify numerically, be very precise (see for example Fig. 2). Our analysis of the optimal c^* will assume a variable $c = Q/L$ that is continuous and unbounded. As noted before, the optimization takes into account the impact of CSI acquisition under TDD.

Let us first focus on deriving the optimal c^* for MF, where $c \in (0,\infty)$ and $\Omega \triangleq \frac{P_t}{P_t+G}$.

Theorem 4. *In the MF-based (G,Q)-vector coded caching with non-negligible CSI costs, the optimal c^* that maximizes $\bar{\mathcal{R}}^{MF}$ in the asymptotic sense is given by the solution to the following expression:*

$$(1 - 2\zeta_{G,Q}c^*)\ln\left(1 + \frac{\Omega}{c^*}\right) - \frac{\Omega(1 - \zeta_{G,Q}c^*)}{\Omega + c^*} = 0. \qquad (23)$$

Proof. We prove the theorem by demonstrating that $\bar{\mathcal{R}}^{\text{MF}}$ as derived in Corollary 2 is concave over $c \in (0,\infty)$. Let us first note that the first derivative of $\bar{R}^{\text{MF}}$ in (10) is given by

$$\frac{\partial \bar{R}^{\text{MF}}}{\partial c} = GL\left[\ln\left(\frac{\Omega + c}{c}\right) + \frac{c}{\Omega + c} - 1\right], \qquad (24)$$

whereas the second derivative is then given by

$$\frac{\partial^2 \bar{R}^{\text{MF}}}{\partial c^2} = -GL\frac{\Omega^2}{c(\Omega + c)^2} < 0. \qquad (25)$$

By differentiating $\bar{\mathcal{R}}^{\text{MF}}$ in Corollary 2 with respect to c, we have that

$$\frac{\partial \bar{\mathcal{R}}^{\text{MF}}}{\partial c} = (1 - \zeta_{G,Q}c)\frac{\partial \bar{R}^{\text{MF}}}{\partial c} - \zeta_{G,Q}\bar{R}^{\text{MF}} \qquad (26)$$

$$\frac{\partial^2 \bar{\mathcal{R}}^{\text{MF}}}{\partial c^2} = (1 - \zeta_{G,Q}c)\frac{\partial^2 \bar{R}^{\text{MF}}}{\partial c^2} - 2\zeta_{G,Q}\frac{\partial \bar{R}^{\text{MF}}}{\partial c}. \qquad (27)$$

Let us know to inspect the signs of these derivatives. First, note that $\frac{\Omega + c}{c} \geq 1$ for any feasible Ω, c, simply because $\Omega = \frac{P_t}{P_t+G} \geq 0$. Let us also note that the function $\ln(x) + 1/x$ is decreasing when $x \in (0,1)$ and is increasing when $x \in [1,\infty)$,

[8]The derivation of the optimal point for RZF is omitted due to the fact that, although we can obtain the derivative of the sum-rate, the equation to find the optimal Q provides little insight and we would need to obtain the solution numerically (cf. Appendix III in [56]).

and also that its minimum value (attained at $x = 1$) is equal to 1. Consequently, it follows that

$$\frac{\partial \bar{R}^{\mathrm{MF}}}{\partial c} = GL\left[\ln\left(\frac{\Omega + c}{c}\right) + \frac{c}{\Omega + c} - 1\right] \geq 0, \qquad (28)$$

where the inequality is strict unless $\frac{\Omega+c}{c} = 1$ corresponding to $c \to \infty$. Therefore, we conclude that $\bar{R}^{\mathrm{MF}}$ is monotonically increasing over $c \in (0, \infty)$.

From the fact that $\frac{\partial \bar{R}^{\mathrm{MF}}}{\partial c} \geq 0$ (cf. (28)), the fact that $\frac{\partial^2 \bar{R}^{\mathrm{MF}}}{\partial c^2} < 0$ (cf. (25)), and the fact that $1 - \zeta_{G,Q} c \geq 0$, we can conclude that $\frac{\partial^2 \bar{\mathcal{R}}^{\mathrm{MF}}}{\partial c^2} < 0$ in (27). Therefore, $\bar{\mathcal{R}}^{\mathrm{MF}}$ is concave over $c \in (0, \infty)$, and thus the global maximum point of $\bar{\mathcal{R}}^{\mathrm{MF}}$ is at the root c^* of $\frac{\partial \bar{\mathcal{R}}^{\mathrm{MF}}}{\partial c}$. □

Next, we consider ZF-based cache-aided precoding, for which we have the following.

Theorem 5. *In the ZF-based (G, Q)-vector coded caching with non-negligible CSI costs, the optimal c^* that maximizes $\bar{\mathcal{R}}^{ZF}$ is given by the solution to the following equation:*

$$(1 - 2\zeta_{G,Q} c^*) \ln\big(1 + \tfrac{P_t}{G}\big(\tfrac{1}{c^*} - 1\big)\big) - \frac{(1 - \zeta_{G,Q} c^*)\frac{P_t}{G}}{(1 - \frac{P_t}{G})c^* + \frac{P_t}{G}} = 0. \qquad (29)$$

Proof. The proof builds on the properties of the first and second derivatives of $\bar{\mathcal{R}}^{\mathrm{ZF}}$, in a similar manner as in the proof of Theorem 4. These derivatives now take the form $\frac{\partial \bar{\mathcal{R}}^{\mathrm{ZF}}}{\partial c} = (1 - \zeta_{G,Q} c)\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c} - \zeta_{G,Q} \bar{R}^{\mathrm{ZF}}$, and $\frac{\partial^2 \bar{\mathcal{R}}^{\mathrm{ZF}}}{\partial c^2} = (1 - \zeta_{G,Q} c)\frac{\partial^2 \bar{R}^{\mathrm{ZF}}}{\partial c^2} - 2\zeta_{G,Q}\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c}$. After applying (15), these derivatives take the form

$$\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c} = GL\left[\ln\big(1 + \tfrac{P_t}{G}\big(\tfrac{1}{c} - 1\big)\big) - \frac{\frac{P_t}{G}}{(1 - \frac{P_t}{G})c + \frac{P_t}{G}}\right] \qquad (30)$$

$$\frac{\partial^2 \bar{R}^{\mathrm{ZF}}}{\partial c^2} = -GL\,\frac{(\frac{P_t}{G})^2}{c\big(\big(1 - \frac{P_t}{G}\big)c + \frac{P_t}{G}\big)^2} < 0. \qquad (31)$$

Since the second derivative $\frac{\partial^2 \bar{R}^{\mathrm{ZF}}}{\partial c^2}$ in (31) is always negative, $\bar{R}^{\mathrm{ZF}}$ is a concave function with respect to c. Therefore, the root of $\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c} = 0$, which we denote by $c_R^\star$, is the global maximum of $\bar{R}^{\mathrm{ZF}}$ over $c \in (0, \infty)$. Moreover, it follows from (15) that $\bar{R}^{\mathrm{ZF}} = 0$ for $c = 1$ and that $\bar{R}^{\mathrm{ZF}} > 0$ for $0 < c < 1$, which implies that $c_R^\star$ belongs in the interval $(0, 1)$.

Since $\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c}\big|_{c=c_R^\star} = 0$ and since $\frac{\partial^2 \bar{R}^{\mathrm{ZF}}}{\partial c^2}$ is always negative, we know that $\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c}$ is monotonically decreasing and that this same $\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c}$ is negative for all $c \in (c_R^\star, 1)$.

Consequently, $\bar{\mathcal{R}}^{\mathrm{ZF}}$ is monotonically decreasing in the interval $c \in (c_R^\star, 1)$. Thus the maximum point of $\bar{\mathcal{R}}^{\mathrm{ZF}}$ must belong in the interval $(0, c_R^\star)$ where we can see that $\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c} > 0$ and $\frac{\partial^2 \bar{R}^{\mathrm{ZF}}}{\partial c^2} < 0$. Hence, $\bar{\mathcal{R}}^{\mathrm{ZF}}$ is concave throughout $c \in (0, c_R^\star)$, and thus the root of $\frac{\partial \bar{\mathcal{R}}^{\mathrm{ZF}}}{\partial c}$ is the global maximum point of $\bar{\mathcal{R}}^{\mathrm{ZF}}$, where this point c^* must belong in $(0, c_R^\star)$. Finally, substituting (15) and (30) into $\frac{\partial \bar{\mathcal{R}}^{\mathrm{ZF}}}{\partial c}$ yields (29) and proves the theorem. □

Remark 2. *As $P_t \to \infty$, we can write* (30) *as* $\frac{\partial \bar{R}^{\mathrm{ZF}}}{\partial c} = GL\big[\ln\big(\frac{P_t}{G}\big) + \ln\big(\frac{1-c}{c}\big) - \frac{1}{1-c}\big] + o(1)$, *where* $\lim_{P_t \to \infty} o(1) = 0$. *Therefore, in the high-SNR regime and without taking CSI costs into account, the optimal value of c that maximizes $\bar{R}^{ZF}$ (and thus $\bar{R}^{RZF}$, since both converge at high-SNR) is given by $c^* = \big(1 + \frac{1}{\mathcal{W}(P_t/(eG))}\big)^{-1}$, upon omitting an $o(1)$ additive term, and upon using $\mathcal{W}(\cdot)$ to denote the Lambert W-Function. This expression can serve as a good approximation in those moderate-to-high SNR scenarios where the dimensionality of the problem implies a relatively small CSI cost. As one can see, as the SNR becomes very large, the above c^* converges, as is known, to 1, corresponding to $Q \approx L$.*

After deriving the above optimal c^*, we can now consider the ratio

$$\mathcal{G}_i^\star \triangleq \frac{\max_{Q \in \mathbb{Z}^+} \bar{\mathcal{R}}^i(G, Q)}{\max_{Q' \in \mathbb{Z}^+} \bar{\mathcal{R}}^i(G = 1, Q')}, \qquad (32)$$

which describes the performance boost due to caching, over (independently) optimized downlink cacheless systems, after accounting for CSI costs. These gains $\mathcal{G}_{\mathrm{MF}}^\star, \mathcal{G}_{\mathrm{ZF}}^\star, \mathcal{G}_{\mathrm{RZF}}^\star$ are reported for the three precoders of interest. As one would expect, this comparison is done under a fixed set of SNR and antenna resources. The transition from the continuous c to the operating Q will follow by simply considering $Q^* = \arg\max_{Q \in \{\lfloor c^* L \rfloor, \lfloor c^* L \rfloor + 1\}} \{\bar{\mathcal{R}}(Q)\}$, where $\lfloor \cdot \rfloor$ denotes the nearest integer less than or equal to the argument.

V. Numerical Results

We proceed to numerically demonstrate the achieved effective rates as well as the effective gains that an optimized vector coded caching scheme provides over the independently optimized cacheless downlink solution. We note that the simulated results employ no approximations (for example, the corresponding SINR is taken directly from (6)). For ease of exposition, we list in Table I the derived theorems and corollaries.

The following figures build on the analysis of the effective sum rates and effective gains of Section III, as well as on the analysis of the optimized gains of Section IV. These figures incorporate the CSI costs in the realistic scenario of having $\beta_{\mathrm{tot}} = 10$, $T_c = 0.04$ seconds and $W_c = 300$ kHz (cf. (22)), which captures the common scenario of low-mobility users consuming videos. We note that $\beta_{\mathrm{tot}} = 10$ is high enough to allow us to neglect the impact of CSI estimation noise [54], [55]. Fig. 2 (left) describes the effective rate of the different cache-aided schemes, for different values of Q. The plot highlights the tightness of the results of Theorems 1–3 (after accounting for CSI costs: see Corollary 2), where we see that indeed the derived asymptotically-approximate expressions have no discernible distance from the actual (simulated) performance. The vertical lines indicate the optimal Q derived in Theorems 4–5. These optimal points indeed match the actual maximum point of the curves. Fig. 2 (right) extends this illustration of the tightness of the results, to Theorems 4–5, by illustrating the optimized (over all Q choices) effective rate performance of the three precoders, comparing the derived results to the actual performance. We note that, for the case of RZF precoding, we represent the result of Theorem 3 (Corollary 2) by considering a c^* value that is obtained from an exhaustive search based on these derived expressions.

TABLE I: Derived Theorems (Thms.) and Corollaries (Cors.)

Thm. 1	Thm. 2	Thm. 3	Thm. 4	Thm. 5	Cor. 1	Cor. 2	Cor. 3
Average sum-rate MF	Average sum-rate ZF	Average sum-rate RZF	Optimal Q for MF	Optimal Q for ZF	Average sum-rate in cacheless MF	Effective rates in MF/ZF/RZF	Effective gains in MF/ZF/RZF

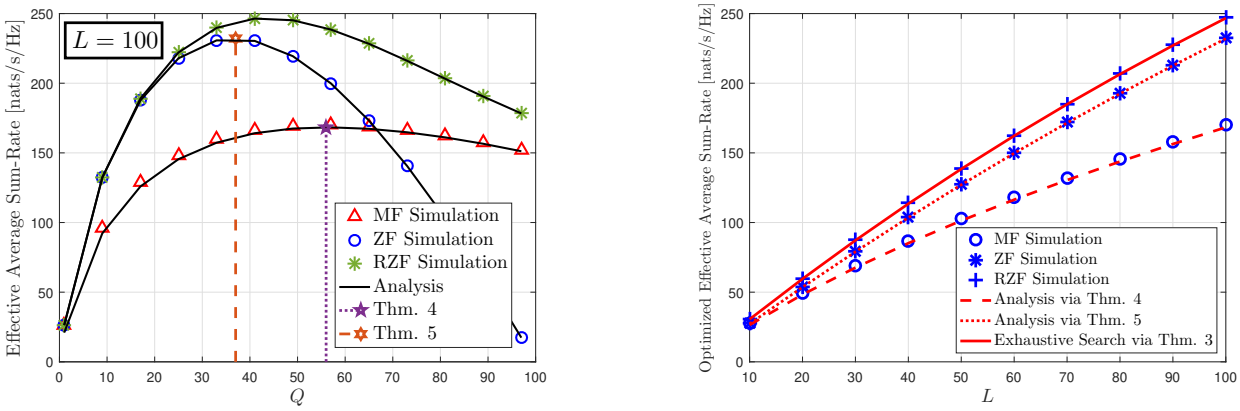


Fig. 2: Effective rate $\bar{\mathcal{R}}$ and optimized effective rate for $P_t = 10$ dB and $G = 5$.

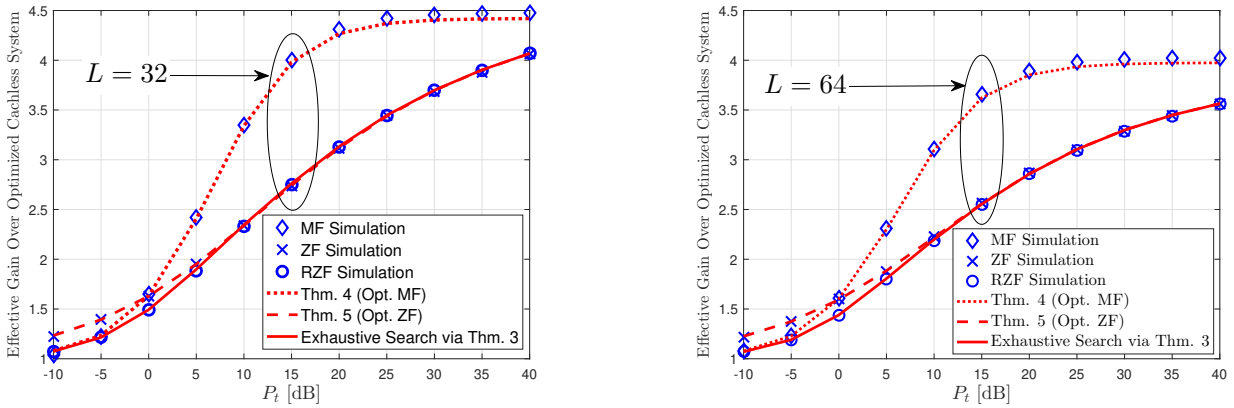


Fig. 3: Effective gain $\mathcal{G}^\star$ over optimized cacheless system for $L \in \{32, 64\}$ and $G = 6$.

Fig. 3 focuses on the effective gains over optimized cacheless downlink systems. As before, the theoretical and simulated results fully match. Here the theoretical results reflect the effective gain ratio $\mathcal{G}_i^\star$ in (32), where the derived effective-rate expressions are from Corollary 2 (and the corresponding Theorems 1–3), and where the optimized c^* are directly from[9] Theorems 4–5.

It is notable that, despite the fact that Theorem 1 and Theorem 3 are obtained from asymptotic analysis, they closely characterize the real performance obtained from simulations. This is also reflected in Fig. 4.

Under the above realistic coherence periods and coherence bandwidths, realistic CSI costs, as well as realistic values of SNR and L, the multiplicative boosts over the achievable rates of optimized downlink systems are quite notable. For example, for 64 transmit antennas, a receiver-side SNR of 20 dB, the same $W_c = 300$ kHz and $T_c = 40$ ms, and under realistic file-size and cache-size constraints that allow us to assume $G = 6$, vector coded caching is here shown to offer a multiplicative boost of about 280% in ZF/RZF precoding and 380% over MF-based cacheless systems, whereas for the case of 32 antennas the gain elevates to 310% for ZF and to a 430% multiplicative boost in the performance of already optimized MF-based cacheless systems[10]. As one would expect, this same figure reveals that the gains $\mathcal{G}^\star_{\text{MF}}, \mathcal{G}^\star_{\text{ZF}}, \mathcal{G}^\star_{\text{RZF}}$ grow monotonically with the SNR, and often come very close to the theoretical upper bound of G.

Another interesting comparison is shown in Fig. 4, where we ask that the cache-aided and cacheless scenarios share the same exact multiplexing gain Q. The motivation for this comparison traces back to the idea of channel hardening, which refers to the fact that as long as L is sufficiently large, and as long as Q/L is sufficiently small, the channel converges to a deterministic value, thus making CSI acquisition easier. While this paper is not about the channel hardening properties of the cache-aided downlink, Fig. 4 — which plots the effective gain $\mathcal{G}(G, Q; 1, Q) = \bar{\mathcal{R}}(G, Q)/\bar{\mathcal{R}}(1, Q)$ — offers a first indication

[9]We recall that, for the RZF case, in Fig. 3 we numerically evaluate c^* from Theorem 3.

[10]In addition to the speedup factor reported here, the use of caches can also lead to additional — albeit marginal — reductions in delivery-time, complements of the so-called local caching gain, which is though of no particular interest to this study.

of yet another benefit of vector coded caching, which now allows us to serve more users at a time, but do so with a controlled ratio Q/L that guarantees certain channel hardening conditions. Focusing on the case of a fixed $Q = 8$ for both the cache-aided ($G = 6$), as well as the cacheless case ($G = 1$), Fig. 4 reveals that under the same W_c, T_c and under realistic SNR values of, for example, approximately 15dB, the effective gains (over cacheless equivalent systems with the same Q/L) approach 400% for the ZF-based precoders, and even go beyond 540% when using MF-based precoding. Similar gains are recorded in the larger scenario with $L = 128$ transmit antennas.

So far, for the sake of clarity of exposition, we have considered the case where K is a multiple of Λ. The impact of deviating from this assumption is indeed very small. Let us briefly discuss this. Let $\underline{\mathrm{B}} \triangleq \lfloor K/\Lambda \rfloor$, in which case $K - \Lambda\underline{\mathrm{B}}$ cache groups will have $\underline{\mathrm{B}} + 1$ users each, while the remaining $\Lambda(\underline{\mathrm{B}} + 1) - K$ cache groups will have $\underline{\mathrm{B}}$ users. Then for the first $\frac{\underline{\mathrm{B}}+1}{Q} - 1$ delivery processes, the effective gain will be the same as before, while for the remaining processes this will be slightly reduced. Let us consider the worst case where there are only $\underline{\mathrm{B}}$ users in each cache group in a specific cache-group set Ψ. In this case, we can have that the number of users in each cache group in Ψ is $\Gamma Q + (Q - 1)$ where $\Gamma = \frac{\underline{\mathrm{B}}+1}{Q} - 1$. The corresponding effective gain averaged over the entire $\Gamma + 1$ delivery processes for serving the cache-group set Ψ is then

$$\mathcal{G}(G, Q; 1, Q') = \frac{1}{\Gamma + 1}\left(\Gamma \frac{\bar{\mathcal{R}}(G, Q)}{\bar{\mathcal{R}}(1, Q')} + \frac{\bar{\mathcal{R}}(G, Q - 1)}{\bar{\mathcal{R}}(1, Q')}\right), \quad (33)$$

where $\bar{\mathcal{R}}(\cdot)$ was introduced in Definition 1. We illustrate this result in Fig. 5, which plots this effective gain in (33), comparing it to the corresponding gain under the assumption that Λ divides K (denoted by $\Lambda | K$ in Fig. 5). We can easily see that the effective gain gap decreases as Γ increases, and eventually becomes negligible for a reasonable value of Γ, e.g., gap $\leq 2\%$ for $\Gamma = 4$ in both the medium and the high SNR regimes.

The above numerical illustrations refer to theoretical gains of $G = 5$ and $G = 6$. To better understand the implications that such values entail, we provide the following simplifying example scenario in which we explain how the considered values are obtained in some realistic use cases.

Example 1. *Let us consider the Netflix library, focusing on movies, and let us make the educated speculation that the popularity distribution of the library content follows a Zipf distribution with exponent parameter* 1.4 *(cf. [57]). Assume that we choose to apply coded caching on the part of the library that captures* 90% *of the traffic, such that on average* 90% *of the Netflix traffic will experience a streaming volume reduction by a (theoretical) factor of* G*. Thus in a Netflix library of approximately* 3700 *movies, coded caching is applied to the* 100 *most popular ones. The remaining 10% of the traffic is sent in an uncoded manner.*

The subpacketization constraint will be largely defined by the latency requirements, which will ask from us, before subpacketization, to first split each movie into files that — in order to guarantee smooth streaming — will have to be sufficiently small. Assume a latency of two minutes, which can be seamlessly handled with a small buffer. This, with the extra assumption that movies last around 90 *minutes, implies file (sub-movie) sizes of approximately* $\frac{2\text{min}}{90\text{min}} = 1/45$ *of the movie size. Let us now consider several possible scenarios that we can encounter in practice.*

a) First setting: Let us assume that the receiving devices are each endowed with a cache of size equal to $25GB$*, and let us assume that they stream HD movies whose size is approximately* $1.3GB$*. From this, we obtain a file size of approximately* $\frac{1.3GB}{45} = 28.8MB$*.*

Under the assumption of atomic (indivisible) communication packets of size equal to 50 *bytes, this brings us to a subpacketization of* $\frac{28.8\text{MB}}{50\text{B}} \approx 6 \cdot 10^5$*. This level of subpacketization, together with the corresponding* $\gamma = \frac{25GB}{100 \cdot 1.3GB} \approx 0.19$*, allows for a theoretical gain of* $G = 7$ *(since* $\binom{\Lambda}{0.19\Lambda} \leq 6 \cdot 10^5$ *and* $G = \Lambda\gamma + 1$*). Recalling our example of the hardening-constrained setting with* $Q = 8$ *(cf. Fig. 4), to attain the promised gain of* $G = 7$*, we require at least* $Q\Lambda \approx 240$ *receiving nodes/antennas, which could represent* 60 *users with* 4 *receive antennas each.*

b) Second setting: Under approximately the same conditions, but for Full-HD movies of size $2.47GB$*, the corresponding scenario implies* $\gamma < 0.10$ *and can allow for a gain close to* $G = 6$*. Recalling the same setting with* $Q = 8$ *of Fig. 4, under the Full-HD assumption, we see that attaining the promised gain of* $G = 6$ *requires a network with at least* $Q\Lambda \approx 400$ *receiving nodes/antennas, which could represent* $K = 100$ *users with* 4 *receive antennas each.*

c) Third setting: Let us now assume that each cache has a size equal to $5GB$*, and let us consider Standard Definition (SD-*$480p$*) streaming. Hence, the file (sub-movie) sizes become* $\frac{400.5MB}{45} = 8.9MB$ *and* $\gamma = \frac{5GB}{100 \cdot 400.5MB} \approx 0.125$*. With an atomic communication packet size of* 200 *bytes, we have subpacketization* $4.5 \cdot 10^4$*, with a theoretical gain of* $G = 5$*. In this SD small-cache scenario, considering* $Q = 8$ *corresponds to* $K = 256$ *single antenna users, or* 128 *users with* 2 *antennas each.*

VI. Conclusions and Future Perspectives

This work explores new methods for improving the performance of advanced multi-antenna downlink systems. Such systems constitute the backbone of modern wireless communications, and they have traditionally depended on an optimized interplay between multiplexing and beamforming techniques. While multi-antenna arrays have been without a doubt one of the most valuable resources and the driving force behind advanced communications technologies, we are now presented with a new and highly complementary and abundant resource in the form of the ever-increasing storage volumes available across even the smallest communicating nodes.

Motivated by the opportunity offered by this newly abundant resource, our work presented very simple to implement optimized cache-aided linear precoding schemes for the multi-antenna downlink broadcast channel. These schemes simply exploit cached content in order to be able to simultaneously transmit carefully selected precoded vectors that would have otherwise been sent one after the other. Because of the simplicity of this idea, it is conceivable to expect the gains to persist

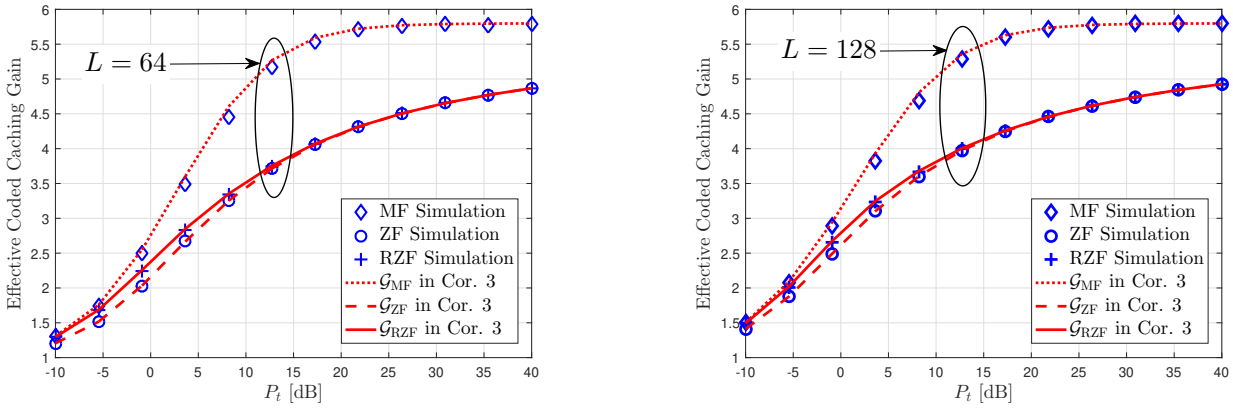


Fig. 4: Hardening-constrained effective gain over a constrained classical downlink system. Q is fixed at $Q = 8$, while $G = 6$.

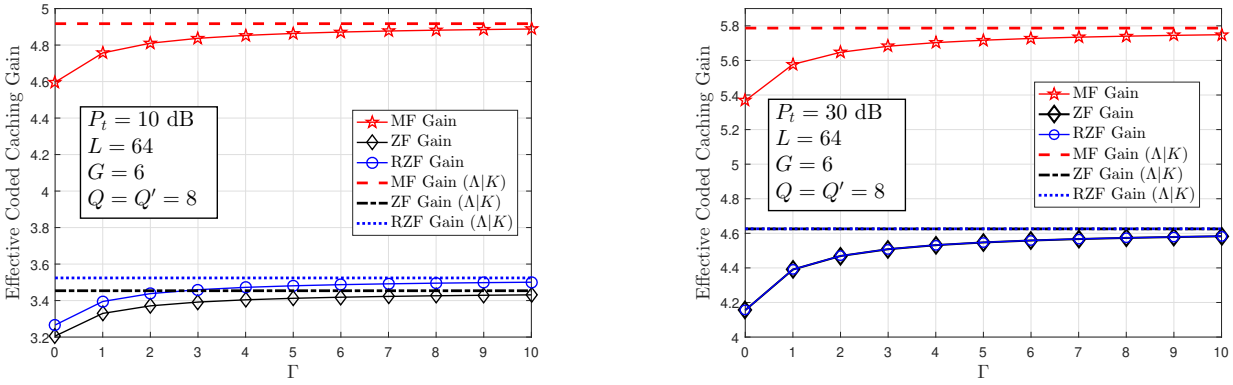


Fig. 5: Effective gain versus Γ in medium SNR (10 dB) and high SNR (30 dB).

for a broader class of precoders. Our performance analysis derives simple expressions that reveal significant multiplicative gains from applying caching over already optimized downlink systems, where these gains persist for various well-known precoding classes. This same analysis and optimization are here shown to hold very tight in realistic non-asymptotic settings, while also incorporating a variety of practical considerations such as power dissemination across signals, realistic SNR values, as well as CSI costs. The comparisons of optimized cache-aided vs. optimized cacheless downlink systems reveal that vector coded caching can recover a sizeable portion of its theoretic (high-SNR) gain $G = \Lambda\gamma + 1$, even in realistic wireless settings operating at realistic SNR values.

In terms of challenges, indeed G remains, under current practices, bounded in the range of single digits. Any improvement beyond this range would require either a dramatic increase in the storage capability of nodes (γ), or a research breakthrough in the area of subpacketization-constrained coded caching. Further improving the subpacketization-constrained performance of coded caching primitives (thus effectively allowing for a larger Λ) remains to date the big challenge in coded caching, and any progress in that direction would undoubtedly have a profound impact on the performance of cache-aided multi-antenna systems.

The reported gains here will naturally come under pressure from additional realistic considerations such as having statistically asymmetric channels, although this problem can be partially ameliorated with power control, with rate-splitting approaches [58], [59], or with the novel "brothers" approach in [60]. These same reported gains may also come under pressure from the additional CSI costs that would arise in the event where multi-antenna coded caching algorithms start serving more and more users. Remedies for this can be found in the novel clique structures recently reported in [51]. A big associated open problem is the simultaneous reduction of both the subpacketization and CSI costs (see [61] for some early efforts). Naturally the system performance also remains subject to the need for cacheable and live-streamed data to co-exist (cf. [59]), the need for cache-aided and cacheless users to coexist,[11] as well as will depend on the stochastic nature of the network topology and user behavior (for some early remedies, the reader can refer to [40], [62]).

The presented new results, as well as the aforementioned challenges, arrive at an instance when bandwidth and antenna resources are asked to handle an aggressively increasing volume of data. At the same time though, the new results come at a time when Moore's law on storage capabilities remains intact and the ever-increasing majority of communicated content is cacheable [3]. For these reasons, and given the powerful gains reported here, we believe that the aforementioned techniques can further help translate the abundance of Gbytes of storage space into much needed spectral efficiency.

[11] See [30], which reveals the surprising conclusion that cacheless users can benefit from full coded caching gains.

APPENDIX I: PROOF OF THEOREM 1

Similar to the proof of [49, Lemma 1], we define $X \triangleq \frac{P_t}{GcL^2}|\mathbf{h}_{\psi,k}^T\mathbf{h}_{\psi,k}^*|^2$ and $Y \triangleq 1+\frac{1}{Q}\sum_{\vartheta=1,\vartheta\neq k}^{Q} Y_\vartheta$, where $Y_\vartheta \triangleq \frac{P_t}{GL}|\mathbf{h}_{\psi,k}^T\mathbf{h}_{\psi,\vartheta}^*|^2$. From [48, Lemma 1], we know that $\mathbb{E}\{X\} = \frac{P_t}{cG}(1+\frac{1}{L})$, $\mathrm{Var}\{X\} = \frac{P_t^2}{G^2c^2}\left(\frac{4}{L}+\frac{10}{L^2}+\frac{6}{L^3}\right) < \infty$, $\mathbb{E}\{Y_\vartheta\} = \frac{P_t}{G}$ and $\mathrm{Var}\{Y_\vartheta\} = \frac{P_t^2}{G^2}(1+\frac{2}{L}) < \infty$. We want to prove that

$$\begin{aligned}\frac{\bar{R}^{\mathrm{MF}}(G,cL)}{c\,GL} &= \mathbb{E}\left\{\ln\left(1+\frac{X}{Y}\right)\right\}\\ &= \ln\left(1+\tfrac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}}\right)+o(1),\ \text{as } Q=cL\to\infty. \end{aligned} \quad (34)$$

By applying Jensen's inequality on $\mathbb{E}\{\ln(X+Y)\}$ and $\mathbb{E}\{\ln(Y)\}$ separately, we get the next two bounding results.

$$\ln\left(\tfrac{1}{\mathbb{E}\{(X+Y)^{-1}\}}\right) \le \mathbb{E}\{\ln(X+Y)\} \le \ln(\mathbb{E}\{X+Y\}) \quad (35)$$

$$-\ln(\mathbb{E}\{Y\}) \le -\mathbb{E}\{\ln(Y)\} \le -\ln\left(\tfrac{1}{\mathbb{E}\{Y^{-1}\}}\right), \quad (36)$$

and after combining these two bounds, we get

$$\begin{aligned}\ln\left(\tfrac{1}{\mathbb{E}\{(X+Y)^{-1}\}}\right) - \ln(\mathbb{E}\{Y\}) &\le \mathbb{E}\left\{\ln\left(1+\tfrac{X}{Y}\right)\right\}\\ &\le \ln(\mathbb{E}\{X+Y\}) - \ln\left(\tfrac{1}{\mathbb{E}\{Y^{-1}\}}\right). \end{aligned} \quad (37)$$

On the other hand, Jensen's inequality says that $\mathbb{E}\{Y^{-1}\} \ge 1/\mathbb{E}\{Y\}$ and $\mathbb{E}\{(X+Y)^{-1}\} \ge 1/\mathbb{E}\{(X+Y)\}$, which yields

$$\ln\left(1+\frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}}\right) \le \ln(\mathbb{E}\{X+Y\}) - \ln\left(\frac{1}{\mathbb{E}\{Y^{-1}\}}\right) \quad (38)$$

$$\ln\left(1+\frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}}\right) \ge \ln\left(\tfrac{1}{\mathbb{E}\{(X+Y)^{-1}\}}\right) - \ln(\mathbb{E}\{Y\}). \quad (39)$$

At this point, both $\mathbb{E}\left\{\ln\left(1+\frac{X}{Y}\right)\right\}$ and $\ln\left(1+\frac{\mathbb{E}\{X\}}{\mathbb{E}\{Y\}}\right)$ are bounded above and below by the same bounds (37)–(39), and the gap between these bounds (Δ) takes the form

$$\begin{aligned}\Delta &\triangleq \left\{\ln(\mathbb{E}\{X+Y\}) - \ln\left(\frac{1}{\mathbb{E}\{Y^{-1}\}}\right)\right\}\\ &\quad - \left\{\ln\left(\frac{1}{\mathbb{E}\{(X+Y)^{-1}\}}\right) - \ln(\mathbb{E}\{Y\})\right\}\\ &= \ln\left[\left(\mathbb{E}\{X+Y\}\,\mathbb{E}\left\{(X+Y)^{-1}\right\}\right)\left(\mathbb{E}\{Y\}\,\mathbb{E}\left\{Y^{-1}\right\}\right)\right].\end{aligned}$$

We want to show that this gap vanishes as $Q = cL \to \infty$. By expanding the Taylor series of Y^{-1} at $\mathbb{E}\{Y\}$, we have that

$$\begin{aligned}&\lim_{Q\to\infty}\mathbb{E}\{Y\}\,\mathbb{E}\left\{Y^{-1}\right\}\\ &= \lim_{Q\to\infty}\mathbb{E}\{Y\}\,\mathbb{E}\left\{\tfrac{1}{\mathbb{E}\{Y\}} - \tfrac{(Y-\mathbb{E}\{Y\})}{\mathbb{E}^2\{Y\}} + \tfrac{(Y-\mathbb{E}\{Y\})^2}{\mathbb{E}^3\{Y\}} + \cdots\right\}\\ &= 1+\lim_{Q\to\infty}\mathbb{E}\{g(Y)\} \overset{(a)}{=} 1+\mathbb{E}\Big\{\lim_{Q\to\infty} g(Y)\Big\} \overset{(b)}{=} 1, \end{aligned} \quad (40)$$

where $g(Y) \triangleq \sum_{n=2}^{\infty}(-1)^n\frac{(Y-\mathbb{E}\{Y\})^n}{\mathbb{E}^n\{Y\}}$, where (a) follows from exchanging the order of the limitation and expectation operators (validated via the Dominated Convergence Theorem (DCT))[12], and where (b) follows from using the DCT to exchange the limitation and infinite summation operators in $\lim_{Q\to\infty} g(Y)$ (similar to the step (a)) and then by considering that $Y - \mathbb{E}\{Y\} \to 0$ as $Q \to \infty$ (due to the law of large numbers). By using similar mathematical manipulations, we have that

$$\lim_{Q=cL\to\infty}\mathbb{E}\{X+Y\}\,\mathbb{E}\left\{(X+Y)^{-1}\right\} = 1. \quad (41)$$

Considering the two limits (40) and (41), we can directly conclude that $\lim_{Q=cL\to\infty}\Delta = 0$, and therefore prove (34).

Finally, substituting $\mathbb{E}\{X\} = \frac{P_t}{cG}(1+\frac{1}{L})$ and $\mathbb{E}\{Y\} = 1+\frac{P_t}{G}\frac{Q-1}{Q}$ into (34) and considering $Q = cL \to \infty$, completes the proof of Theorem 1.

[12]To see this, first define $Z \triangleq |Y - \mathbb{E}\{Y\}| \ge 0$. As $Q \to \infty$, $Z \to 0$ (due to the law of large numbers), there always exists a constant Q_0 and $\varepsilon < 1$ such that $Z < \varepsilon$ for any $Q > Q_0$. For $Z < \varepsilon$, we have that $\sum_{n=2}^{\infty} Z^n = \frac{Z^2}{1-Z} < \frac{\varepsilon^2}{1-\varepsilon}$. Considering $g(Y) \le \sum_{n=2}^{\infty} Z^n$ and $\mathbb{E}\{\sum_{n=2}^{\infty} Z^n\} < \frac{\varepsilon^2}{1-\varepsilon} < \infty$, which satisfies the DCT condition, yields that $\lim_{Q\to\infty}\mathbb{E}\{g(Y)\} = \mathbb{E}\{\lim_{Q\to\infty} g(Y)\}$.

APPENDIX II: PROOF OF THEOREM 3

We split the proof in three parts. First, we present the proof of (19). Then, we provide two useful lemmas, and we conclude by deriving the asymptotic deterministic equivalent of the SINR.

A. Proof of (19)

We provide here the proof of the expression of $\mathrm{SINR}_{\psi,k}^{\mathrm{RZF}}$ in (19). Let us recall that $\mathbf{H}_{\psi,-k}$ represents the matrix $\mathbf{H}_\psi$ after removing its k-th row. The useful signal contribution to the received signal in (16) (omitting the term $\rho_\psi/\sqrt{G}$ for the sake of conciseness) can be written as

$$\begin{aligned}&\mathbf{h}_{\psi,k}^T\left(\alpha\mathbf{I}_L + \mathbf{H}_\psi^H\mathbf{H}_\psi\right)^{-1}\mathbf{h}_{\psi,k}^* s_{\psi,k}\\ &= \mathbf{h}_{\psi,k}^T\Big(\alpha\mathbf{I}_L + \mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k} + \mathbf{h}_{\psi,k}^*\mathbf{h}_{\psi,k}^T\Big)^{-1}\mathbf{h}_{\psi,k}^* s_{\psi,k}\\ &\overset{(a)}{=} \frac{\mathbf{h}_{\psi,k}^T\Big(\alpha\mathbf{I}_L + \mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\Big)^{-1}\mathbf{h}_{\psi,k}^*}{1+\mathbf{h}_{\psi,k}^T\left(\alpha\mathbf{I}_L + \mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-1}\mathbf{h}_{\psi,k}^*} s_{\psi,k}\\ &\overset{(b)}{=} \frac{A_{\psi,k}}{1+A_{\psi,k}} s_{\psi,k}, \end{aligned} \quad (42)$$

where (a) follows from the relation

$$\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1}\mathbf{B}\mathbf{D}^{-1} = \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \quad (43)$$

and where (b) follows after applying the definition of $A_{\psi,k}$ from (17).

On the other hand, the power of the interference averaged over data signals in (16) is given by (44) on the top of the next page. We derive (46) by applying again into (45) the matrix identity from (43) and considering the definitions of $A_{\psi,k}$ and $B_{\psi,k}$ from (17)–(18). Then, combining (46) with (42) yields the expression of $\mathrm{SINR}_{\psi,k}^{\mathrm{RZF}}$ in (19). This concludes the proof.

B. Two Useful Lemmas

In the following, we present two lemmas that are instrumental in the derivation of Theorem 3.

Lemma 1. *For any fixed c, $0 < c < \infty$, the trace of $\frac{1}{L}\left(z\mathbf{I}_L + \frac{1}{L}\mathbf{H}_\psi^H\mathbf{H}_\psi\right)^{-1}$ converges to $S_c(z)$ almost surely as $L \to \infty$, where $S_c(z)$ is defined as*

$$S_c(z) \triangleq \frac{1}{2}\left(\sqrt{\frac{(1-c)^2}{z^2}+\frac{2(1+c)}{z}+1}+\frac{1-c}{z}-1\right). \quad (47)$$

$$\left|I_{\psi,k}\right|^2 = \frac{\rho_\psi^2}{G}\sum\nolimits_{\vartheta=1,\,\vartheta\neq k}^{L}\sum\nolimits_{\vartheta'=1,\,\vartheta'\neq k}^{L}\mathbf{h}_{\psi,\vartheta}^T\left(\alpha\mathbf{I}_L+\mathbf{H}_\psi^H\mathbf{H}_\psi\right)^{-1}\mathbf{h}_{\psi,k}^*\mathbf{h}_{\psi,k}^T\left(\alpha\mathbf{I}_L+\mathbf{H}_\psi^H\mathbf{H}_\psi\right)^{-1}\mathbf{h}_{\psi,\vartheta'}^*\mathbb{E}\{s_{\psi,\vartheta}^* s_{\psi,\vartheta'}\} \quad (44)$$

$$= \frac{\rho_\psi^2}{G}\mathbf{h}_{\psi,k}^T\left(\alpha\mathbf{I}_L+\mathbf{H}_\psi^H\mathbf{H}_\psi\right)^{-1}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\left(\alpha\mathbf{I}_L+\mathbf{H}_\psi^H\mathbf{H}_\psi\right)^{-1}\mathbf{h}_{\psi,k}^* \quad (45)$$

$$= \frac{\rho_\psi^2}{G}\frac{\mathbf{h}_{\psi,k}^T\left(\alpha\mathbf{I}_L+\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-1}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\left(\alpha\mathbf{I}_L+\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-1}\mathbf{h}_{\psi,k}^*}{\left(1+\mathbf{h}_{\psi,k}^T\left(\alpha\mathbf{I}_L+\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-1}\mathbf{h}_{\psi,k}^*\right)^2} = \frac{B_{\psi,k}\rho_\psi^2/G}{(1+A_{\psi,k})^2}. \quad (46)$$

Proof. This lemma can be obtained as a direct application of a known result from [63, Ch. 3] for the Stieltjes transform [64]. Hence, we omit the proof due to the page limitation and refer the reader to [63, Ch. 3] for more details. □

Lemma 2. *For any fixed* $0 < c < \infty$ *and arbitrary* $0 < \theta < \infty$*, we have that, as* $L \to \infty$*,*

$$\mathrm{Tr}\left\{\tfrac{1}{L}\left(\theta\mathbf{I}+\tfrac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-2}\right\} \xrightarrow{a.s.} \mathrm{Tr}\left\{\tfrac{1}{L}\left(\theta\mathbf{I}+\tfrac{1}{L}\mathbf{H}_{\psi}^H\mathbf{H}_{\psi}\right)^{-2}\right\}. \quad (48)$$

Proof. Let us first define $\mathbf{A} \triangleq \theta\mathbf{I}+\frac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}$, and let us also define

$$\delta \triangleq \left|\mathrm{Tr}\left\{\frac{1}{L}\left(\theta\mathbf{I}+\frac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-2}\right\} - \mathrm{Tr}\left\{\frac{1}{L}\left(\theta\mathbf{I}+\frac{1}{L}\mathbf{H}_{\psi}^H\mathbf{H}_{\psi}\right)^{-2}\right\}\right|. \quad (49)$$

By applying the Woodbury matrix identity [65], we have that

$$\delta = \left|\frac{1}{L}\mathrm{Tr}\left\{\frac{2}{L}\frac{\mathbf{h}_k^T\mathbf{A}^{-3}\mathbf{h}_k^*}{1+\frac{1}{L}\mathbf{h}_k^T\mathbf{A}^{-1}\mathbf{h}_k^*} - \frac{1}{L^2}\frac{(\mathbf{h}_k^T\mathbf{A}^{-2}\mathbf{h}_k^*)(\mathbf{h}_k^T\mathbf{A}^{-2}\mathbf{h}_k^*)}{\left(1+\frac{1}{L}\mathbf{h}_k^T\mathbf{A}^{-1}\mathbf{h}_k^*\right)^2}\right\}\right|, \quad (50)$$

which can be further rewritten as $\delta = \left|\Theta_1-\Theta_2\right|$, where $\Theta_1 \triangleq \frac{2}{L^2}\frac{\mathbf{h}_k^T\mathbf{A}^{-3}\mathbf{h}_k^*}{1+\frac{1}{L}\mathbf{h}_k^T\mathbf{A}^{-1}\mathbf{h}_k^*}$, and $\Theta_2 \triangleq \frac{1}{L^3}\left(\frac{\mathbf{h}_k^T\mathbf{A}^{-2}\mathbf{h}_k^*}{1+\frac{1}{L}\mathbf{h}_k^T\mathbf{A}^{-1}\mathbf{h}_k^*}\right)^2$. Furthermore, we apply eigenvalue decomposition by factorizing $\frac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}$ as $\frac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H$, which yields $\mathbf{A}^{-1} = \mathbf{Q}\left(\theta\mathbf{I}+\mathbf{\Lambda}\right)^{-1}\mathbf{Q}^H$, and $\mathbf{A}^{-3} = \mathbf{Q}\left(\theta\mathbf{I}+\mathbf{\Lambda}\right)^{-3}\mathbf{Q}^H$. Thus, upon defining $\mathbf{g} \triangleq \mathbf{Q}\mathbf{h}_k^*/\sqrt{L}$, the term Θ_1 can be rewritten as

$$\Theta_1 = \frac{2}{L^2}\frac{\mathbf{h}_k^T\mathbf{Q}\left(\theta\mathbf{I}+\mathbf{\Lambda}\right)^{-3}\mathbf{Q}^H\mathbf{h}_k^*}{1+\frac{1}{L}\mathbf{h}_k^T\mathbf{Q}\left(\theta\mathbf{I}+\mathbf{\Lambda}\right)^{-1}\mathbf{Q}^H\mathbf{h}_k^*} = \frac{2}{L}\frac{\mathbf{g}^H\left(\theta\mathbf{I}+\mathbf{\Lambda}\right)^{-3}\mathbf{g}}{1+\mathbf{g}^H\left(\theta\mathbf{I}+\mathbf{\Lambda}\right)^{-1}\mathbf{g}} = \frac{2}{L}\frac{\sum_{\ell=1}^L|g_\ell|^2\frac{1}{(\theta+\lambda_\ell)^3}}{1+\sum_{\ell=1}^L|g_\ell|^2\frac{1}{\theta+\lambda_\ell}}$$
$$\leq \frac{2}{\theta^2 L}\frac{\sum_{\ell=1}^L|g_\ell|^2\frac{1}{\theta+\lambda_\ell}}{1+\sum_{\ell=1}^L|g_\ell|^2\frac{1}{\theta+\lambda_\ell}} \leq \frac{2}{\theta^2 L} \xrightarrow{L\to\infty} 0, \quad (51)$$

where g_ℓ is the ℓ-th element of $\mathbf{g}$ and λ_ℓ is the ℓ-th eigenvalue of $\frac{1}{L}\mathbf{H}_{\psi,k}^H\mathbf{H}_{\psi,k}$. Similarly, we have that

$$\Theta_2 = \frac{1}{L}\left(\frac{\mathbf{g}^H(\theta\mathbf{I}+\mathbf{\Lambda})^{-2}\mathbf{g}}{1+\mathbf{g}^H(\theta\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{g}}\right)^2 \leq \frac{1}{\theta^2 L}\left(\frac{\sum_{\ell=1}^L|g_\ell|^2\frac{1}{\theta+\lambda_\ell}}{1+\sum_{\ell=1}^L|g_\ell|^2\frac{1}{\theta+\lambda_\ell}}\right)^2 \leq \frac{1}{\theta^2 L} \xrightarrow{L\to\infty} 0 \quad (52)$$

Finally, from (51), (52), and from the fact that $\delta \leq \left|\Theta_1\right|+\left|\Theta_2\right|$, the difference δ approaches zero almost surely as $L \to \infty$. This concludes the proof of Lemma 2. □

C. Proof of Theorem 3

We obtain Theorem 3 by deriving the asymptotic deterministic equivalent of $\mathrm{SINR}_{\psi,k}$ in (19). For that, we first derive the asymptotic deterministic equivalent of $A_{\psi,k}$ and ρ_ψ^2.

Let us start by considering $A_{\psi,k}$, defined in (17). By means of the Trace Lemma and the Rank-1 Perturbation Lemma from [63], we can obtain that

$$A_{\psi,k} = \mathbf{h}_{\psi,k}^T\left(\alpha\mathbf{I}_L+\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-1}\mathbf{h}_{\psi,k}^* \quad (53)$$
$$\xrightarrow{a.s.} \mathrm{Tr}\left\{\left(\alpha\mathbf{I}_L+\mathbf{H}_\psi^H\mathbf{H}_\psi\right)^{-1}\right\} \quad (54)$$

as $L \to \infty$. From this, we can apply Lemma 1 and the fact that $\alpha = L/P_t$ to obtain the deterministic equivalent of $A_{\psi,k}$, which we denote as $a_{\psi,k}$, and which is given by $a_{\psi,k} = S_c\left(\frac{1}{P_t}\right)$, where $S_c(z) = \frac{1}{2}\left[\sqrt{\frac{(1-c)^2}{z^2}+\frac{2(1+c)}{z}+1}+\frac{1-c}{z}-1\right]$ as defined in (47). This yields the expression of $a_{\psi,k}$ in (21).

Next, we focus on $B_{\psi,k}$, introduced in (18), and we again apply the Trace Lemma and the Rank-1 Perturbation Lemma from [63] in the limit of $L \to \infty$ to obtain that

$$B_{\psi,k} \xrightarrow{a.s.} \tfrac{1}{L}\mathrm{Tr}\left\{\tfrac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\left(\tfrac{1}{P_t}\mathbf{I}_L+\frac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-2}\right\}$$
$$= \frac{1}{L}\mathrm{Tr}\left\{\left(\frac{1}{P_t}\mathbf{I}_L+\frac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-1}\right\} - \frac{1}{P_t L}\mathrm{Tr}\left\{\left(\frac{1}{P_t}\mathbf{I}_L+\frac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\right)^{-2}\right\}. \quad (55)$$

The first trace term of the R.H.S. of (55) matches (53), and thus its deterministic equivalent is $a_{\psi,k}$. With respect to the second term of the R.H.S. of (55), applying Lemmas 1 and 2 yields (56) as $L \to \infty$, where $\{\lambda_\ell\}_{\ell=1}^L$ are the eigenvalues of $\frac{1}{L}\mathbf{H}_\psi^H\mathbf{H}_\psi$ and where $\frac{\partial S_c(z)}{\partial z}$ is the derivative of $S_c(z)$ with respect to z, which is given by

$$\frac{\partial S_c(z)}{\partial z} = \frac{1}{2}\left[\frac{-c^2-c(z-2)-z-1}{z^2\sqrt{c^2+2c(z-1)+(z+1)^2}}-\frac{1-c}{z^2}\right]. \quad (57)$$

From (55) and (56) it holds that $B_{\psi,k} \xrightarrow{a.s.} b_{\psi,k} \triangleq a_{\psi,k} + \frac{1}{P_t}\frac{\partial S_c(z)}{\partial z}\Big|_{z=1/P_t}$ as $L \to \infty$.

Finally, we focus on the power control factor for the RZF precoder, which was given by $\rho_\psi^2 = \frac{P_t}{\frac{1}{L}\mathrm{Tr}\{\frac{1}{L}\mathbf{H}_\psi^H\mathbf{H}_\psi(\frac{1}{L}\mathbf{H}_\psi^H\mathbf{H}_\psi+\frac{1}{P_t}\mathbf{I}_L)^{-2}\}}$. From the derivation of $B_{\psi,k}$ and (55)–(56), it follows that $\rho_\psi^2 \xrightarrow{a.s.} \frac{P_t}{b_{\psi,k}}$. Thus, the

$$\frac{1}{P_tL}\mathrm{Tr}\Big\{\Big(\frac{1}{P_t}\mathbf{I}_L+\frac{1}{L}\mathbf{H}_{\psi,-k}^H\mathbf{H}_{\psi,-k}\Big)^{-2}\Big\}\xrightarrow{a.s.}\frac{1}{P_tL}\mathrm{Tr}\Big\{\Big(\frac{1}{P_t}\mathbf{I}_L+\frac{1}{L}\mathbf{H}_{\psi}^H\mathbf{H}_{\psi}\Big)^{-2}\Big\}=\frac{1}{P_tL}\sum\nolimits_{\ell=1}^{L}\frac{1}{(\lambda_\ell+1/P_t)^2}$$
$$=-\frac{1}{P_t}\frac{\partial}{\partial z}\Big(\frac{1}{L}\sum\nolimits_{\ell=1}^{L}\frac{1}{\lambda_\ell+z}\Big)\Big|_{z=1/P_t}=-\frac{1}{P_t}\frac{\partial}{\partial z}\Big(\mathrm{Tr}\Big\{\frac{1}{L}\Big(z\mathbf{I}_L+\frac{1}{L}\mathbf{H}_{\psi}^H\mathbf{H}_{\psi}\Big)^{-1}\Big\}\Big)\Big|_{z=1/P_t}\xrightarrow{a.s.}-\frac{1}{P_t}\frac{\partial S_c(z)}{\partial z}\Big|_{z=1/P_t}. \quad (56)$$

asymptotic deterministic equivalent of ρ_ψ^2, denoted by p_ψ^2, takes the form $p_\psi^2=\frac{P_t}{b_{\psi,k}}$, which, upon substituting (57) in $b_{\psi,k}$, yields the expression of p_ψ^2 in (21).

Next, we obtain the asymptotic deterministic equivalent of $\mathrm{SINR}_{\psi,k}$ by substituting the asymptotic deterministic equivalent of $A_{\psi,k}$, $B_{\psi,k}$ and ρ_ψ^2 into (19), which yields

$$\mathrm{SINR}_{\psi,k}^{\mathrm{RZF}}\xrightarrow{a.s.}\frac{a_{\psi,k}^2p_\psi^2/G}{\big(1+a_{\psi,k}\big)^2+\frac{P_t}{G}}. \quad (58)$$

Finally, a direct application of the Continuous Mapping Theorem [66] yields (20), which concludes the proof of Theorem 3. □

REFERENCES

[1] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Vector coded caching greatly enhances massive MIMO," in *Proc. 23rd IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2022, pp. 1–5.

[2] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[3] "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," White Paper, Cisco, San Jose, CA, USA, Feb. 2019.

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

[5] 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.

[6] 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.

[7] 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, 2018.

[8] H. H. S. Chittoor, P. Krishnan, K. V. S. Sree, and B. Mamillapalli, "Subexponential and linear subpacketization coded caching via projective geometry," *IEEE Trans. Inf. Theory*, vol. 67, no. 9, pp. 6193–6222, Sep. 2021.

[9] E. Parrinello, A. Ünsal, and P. Elia, "Fundamental limits of coded caching with multiple antennas, shared caches and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2252–2268, Apr. 2020.

[10] S. Jin, Y. Cui, H. Liu, and G. Caire, "A new order-optimal decentralized coded caching scheme with good performance in the finite file size regime," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5297–5310, Aug. 2019.

[11] 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.

[12] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Coded placement for systems with shared caches," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[13] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7253–7271, Dec. 2016.

[14] 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.

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

[16] I. Bergel and S. Mohajer, "Cache-aided communications with multiple antennas at finite SNR," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1682–1691, Aug. 2018.

[17] 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.

[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] X. Xu and M. Tao, "Modeling, analysis, and optimization of caching in multi-antenna small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5454–5469, Nov. 2019.

[20] M. Bayat, R. K. Mungara, and G. Caire, "Achieving spatial scalability for coded caching via coded multipoint multicasting," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 227–240, Jan. 2019.

[21] M. J. Salehi, E. Parrinello, S. P. Shariatpanahi, P. Elia, and A. Tölli, "Low-complexity high-performance cyclic caching for large MISO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3263–3278, May 2022.

[22] S. Mohajer and I. Bergel, "MISO Cache-Aided Communication with Reduced Subpacketization," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020.

[23] N. S. Karat, S. Dey, A. Thomas, and B. S. Rajan, "An optimal linear error correcting delivery scheme for coded caching with shared caches," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2019, pp. 1217–1221.

[24] 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)*, 2019, pp. 1–6.

[25] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: Interplay of coded-caching and CSIT feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3142–3160, May 2017.

[26] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[27] T. X. Vu, S. Chatzinotas, and B. Ottersten, "Edge-caching wireless networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2827–2839, Apr. 2018.

[28] M. Salehi and A. Tölli, "Diagonal multi-antenna coded caching for reduced subpacketization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020.

[29] M. Salehi, A. Tölli, and S. P. Shariatpanahi, "Subpacketization-beamformer interaction in multi-antenna coded caching," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[30] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7635–7651, Dec. 2020.

[31] "The industry's first independent benchmark study of 5G NR MU-MIMO," Signals Research Group, Tech. Rep., Sept. 2020.

[32] 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.

[33] H. Zhao, E. Lampiris, G. Caire, and P. Elia, "Multi-antenna coded caching analysis in finite SNR and finite subpacketization," in *Proc. 25th Int. ITG Workshop on Smart Antennas (WSA)*, Nov. 2021, pp. 433–438.

[34] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[35] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[36] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 574–590, Jan. 2018.

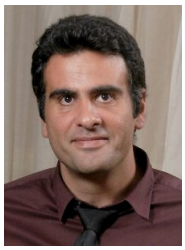
[37] N. Rajatheva *et al.*, "White paper on broadband connectivity in 6G," 2020. [Online]. Available: arxiv.org/abs/2004.14247

[38] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[39] H. Q. Ngo and E. G. Larsson, "No downlink pilots are needed in TDD massive mimo," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2921–2935, May 2017.
[40] A. Malik, B. Serbetci, E. Parrinello, and P. Elia, "Fundamental limits of stochastic shared-cache networks," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4433–4447, Jul. 2021.
[41] C. B. Peel, B. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.
[42] M. Vu and A. Paulraj, "MIMO wireless linear precoding," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 86–105, Sep. 2007.
[43] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, Jul. 2012.
[44] J. Hoydis, S. Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.
[45] M. Matthaiou, M. R. McKay, P. J. Smith, and J. A. Nossek, "On the condition number distribution of complex wishart matrices," *IEEE Trans. Commun.*, vol. 58, no. 6, pp. 1705–1717, Jun. 2010.
[46] C. Feng, Y. Jing, and S. Jin, "Interference and outage probability analysis for massive MIMO downlink with MF precoding," *IEEE Signal Process. Lett.*, vol. 23, no. 3, pp. 366–370, Mar. 2016.
[47] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.
[48] Y.-G. Lim, C.-B. Chae, and G. Caire, "Performance analysis of massive MIMO for cell-boundary users," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6827–6842, Dec. 2015.
[49] Q. Zhang, S. Jin, K.-K. Wong, H. Zhu, and M. Matthaiou, "Power scaling of uplink massive MIMO systems with arbitrary-rank channel means," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 966–981, Oct. 2014.
[50] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations Trends Commun. Inf. Theory*, vol. 1, no. 1, pp. 1–182, Jun. 2004.
[51] E. Lampiris, A. Bazco-Nogueras, and P. Elia, "Resolving the feedback bottleneck of multi-antenna coded caching," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2331–2348, Apr. 2022.
[52] M. Kobayashi and G. Caire, "On the net DoF comparison between ZF and MAT over time-varying MISO broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2012, pp. 2286–2290.
[53] 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.
[54] M. Kobayashi, G. Caire, and N. Jindal, "How much training and feedback are needed in MIMO broadcast channels?" in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2008, pp. 2663–2667.
[55] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jun. 2010.
[56] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Vector coded caching multiplicatively boosts the throughput of realistic downlink systems," 2022. [Online]. Available: https://arxiv.org/abs/2202.07047
[57] S. Gupta and S. Moharir, "Request patterns and caching for VoD services with recommendation systems," in *Proc. Int. Conf. on Commun. Syst. and Netw. (COMSNETS)*, Jan. 2017, pp. 31–38.
[58] E. Lampiris, J. Zhang, O. Simeone, and P. Elia, "Fundamental limits of wireless caching under uneven-capacity channels," in *Proc. Int. Zurich Seminar on Inf. and Commun. (IZS)*, Feb. 2020, pp. 120–124.
[59] H. Joudeh, E. Lampiris, P. Elia, and G. Caire, "Fundamental limits of wireless caching under mixed cacheable and uncacheable traffic," *IEEE Trans. Inf. Theory*, vol. 67, no. 7, pp. 4747–4767, Jul. 2021.
[60] H. Zhao, A. Bazco-Nogueras and P. Elia, "Wireless coded caching can overcome the worst-user bottleneck by exploiting finite file sizes," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5450–5466, Jul. 2022.
[61] 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.
[62] A. Malik, B. Serbetci, and P. Elia, "Coded caching in networks with heterogeneous user activity," Jan. 2022. [Online]. Available: https://arxiv.org/abs/2103.09156
[63] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.
[64] D. V. Widder, "The Stieltjes transform," *Trans. American Mathematical Society*, vol. 43, no. 1, pp. 7–60, 1938.
[65] M. A. Woodbury, "Inverting modified matrices," *Statistical Research Group Memo. Reports, Princeton Univ. (42)*, 1950.
[66] A. W. van der Vaart, *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 2000.

Hui Zhao (Graduate Student Member, IEEE) received the B.S. degree in telecommunications engineering from Southwest University, Chongqing, China, in 2016, and the M.S. degree in electrical engineering from the King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, in 2019. He is currently pursuing the Ph.D. degree in communication systems with EURECOM, Sophia Antipolis, France. His current research interest includes the modeling, design, and performance analysis of wireless communication systems.

Antonio Bazco-Nogueras (Member, IEEE) received the B.S. and M.S. degrees in Telecommunications Engineering from University of Zaragoza, Spain, in 2014 and 2016, respectively. He obtained the Ph.D. degree from Sorbonne Université, Paris, France, in collaboration with the Mitsubishi Electric R&D Centre Europe, Rennes, France, in 2019. He was a postdoctoral researcher at EURECOM, Sophia-Antipolis, France, from 2020 to 2021. He is currently a postdoctoral researcher at IMDEA Networks Institute, Madrid, Spain. He is recipient of the Madrid Talent Attraction Grant 2021. His research interests include multi-user information theory, intelligent and self-configuring networks, decentralized systems, content delivery networks, and cooperative wireless networks.



Petros Elia (Member, IEEE) received the B.Sc. degree from the Illinois Institute of Technology, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, in 2001 and 2006 respectively. He is now a professor with the Department of Communication Systems at EURECOM in Sophia Antipolis, France. His latest research deals with the intersection of coded caching and feedback-aided communications in multiuser settings. He has also worked in the area of complexity-constrained communications, MIMO, queueing theory and cross-layer design, coding theory, information theoretic limits in cooperative communications, and surveillance networks. He is a Fulbright scholar, the co-recipient of the NEWCOM++ distinguished achievement award 2008-2011 for a sequence of publications on the topic of complexity in wireless communications, the recipient of the ERC Consolidator Grant 2017-2022 on cache-aided wireless communications, and the recipient of the ERC-PoC 2022-2024.