Cache-Aided Massive MIMO: Linear Precoding Design and Performance Analysis

Xiao Wei*‡, Lin Xiang§, Laura Cottatellucci‡, Tao Jiang*, and Robert Schober‡

*Huazhong University of Science and Technology, P.R. China; §University of Luxembourg, Luxembourg; ‡Friedrich-Alexander-University Erlangen-Nürnberg, Germany.

Abstract—In this paper, we propose a novel joint caching and massive multiple-input multiple-output (MIMO) transmission scheme, referred to as cache-aided massive MIMO, for advanced downlink cellular communications. In addition to reaping the conventional advantages of caching and massive MIMO, the proposed scheme also exploits the side information provided by cached files for interference cancellation at the receivers. This interference cancellation increases the degrees of freedom available for precoding design. In addition, the power freed by the cache-enabled offloading can benefit the transmissions to the users requesting non-cached files. The resulting performance gains are not possible if caching and massive MIMO are designed separately. We analyze the performance of cache-aided massive MIMO for cache-dependent maximum-ratio transmission (MRT), zero-forcing (ZF) precoding, and regularized zero-forcing (RZF) precoding. Lower bounds on the ergodic achievable rates are derived in closed form for MRT and ZF precoding. The ergodic achievable rate of RZF precoding is obtained for the case when the numbers of transmit antennas and users are large but their ratio is fixed. Compared to *conventional* massive MIMO, the proposed cache-aided massive MIMO scheme achieves a significantly higher ergodic rate especially when the number of users approaches the number of transmit antennas.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) is a key technology to improve the spectral efficiency of cellular communications and thus, to support the explosive growth of cellular traffic [1], [2]. By employing a large number of antennas at the base station (BS), massive MIMO offers abundant spatial degrees of freedom and facilitates large multiplexing and diversity gains.

For a single-cell massive MIMO system, where the BS is equipped with M antennas and communicates with K single-antenna users, performance has been shown to depend critically on the number of BS antennas per user, denoted as $\rho_0 \triangleq M/K$ [3]. For example, when both M and K grow without bound while ρ_0 is finite, the authors of [4] show that the effective signal-to-interference-plus-noise ratio (SINR) grows linearly with ρ_0. In [5], the authors show that simple linear precoders and detectors are asymptotically optimal, i.e., capacity-achieving, in the large system limit where M grows unbounded while K is fixed, i.e., ρ_0 becomes very large. Since the publication of [4] and [5], massive MIMO has been typically studied for low-complexity linear precoding [6], which performs well for large ρ_0. However, due to emerging applications, related to e.g. smart phones, tablets, and Internet-of-Things devices, in future wireless systems, the number of users may grow significantly, even beyond the number of BS antennas [7]. In this case, conventional linear precoding based massive MIMO suffers from a significant performance loss due to the resulting small ρ_0. Hence, improving the performance of massive MIMO

systems when ρ_0 is small is important for future applications but has not been sufficiently addressed in the literature [8].

In this paper, we show that wireless caching at the user side provides an opportunity for enhancing the capacity of massive MIMO, especially for small ρ_0. With the wide-spread use of smart phones and tablets, cache memory is often available at the users. By pre-storing the most popular files in the users' caches during periods of low network traffic, fast access to these files is enabled without requiring over-the-air delivery [9]. However, as the actual users' requests are not known during cache placement, the cached files may not be requested by the users later on, which severely limits the performance gains of user-side caching. To mitigate this problem, two approaches, which exploit additional performance gains enabled by caching, have been proposed in the literature [10], [11]. One approach, referred to as cache-aided non-orthogonal multiple access (NOMA) [10], exploits a user's cached but non-requested files for canceling NOMA interference. By joint optimization of cache-enabled interference cancellation and successive interference cancellation, cache-aided NOMA can significantly improve the users' achievable rates [10]. However, when the number of users is large, the optimization of cache-aided NOMA becomes intractable. An alternative approach employs coded caching [11]. By carefully encoding the cached and delivered files, simultaneous multicast to multiple users is enabled such that each user can decode its requested file without suffering from multiuser interference [11]. However, for coded caching, forming multicast groups and decoding impose a large computational burden on both the BS and the users [11]. Synergies between caching and massive MIMO are explored in [12], where several communication schemes are proposed and analyzed. One of the schemes in [12] combines coded caching with massive MIMO for improved multicast transmission over fading channels, whereas another scheme leverages the spatial multiplexing capability of massive MIMO to simultaneously transmit the requested files to all the users. Finally, a combination of the two schemes is also analyzed in [12].

In this paper, we propose several novel cache-aided precoding schemes for massive MIMO, which are collectively referred to as cache-aided massive MIMO. Assume that each user is equipped with a cache memory and the BS is equipped with a large number of antennas. If the files cached at one user are requested by the user itself, caching offloads the transmission to the user and, by cache-aided massive MIMO, more transmit power can be allocated to the other users. On the other hand, if the files cached at a user are not requested by the user itself but are requested by other users, these files can be exploited

for interference cancellation at the user, avoiding the need for interference suppression via precoding at the BS. Consequently, cache-aided massive MIMO introduces additional degrees of freedom for the transmission of the remaining files which leads to improved performance. Appealingly, owing to the large antenna array at the BS, the performance gains enabled by caching are achievable via properly redesigned linear precoders. Hence, cache-aided massive MIMO avoids the encoding and decoding overhead incurred by coded caching [11] and cache-aided NOMA [9], and is computationally efficient. The main contributions of this paper are as follows:

- We propose a novel cache-aided massive MIMO scheme, which not only facilitates offloading and interference cancellation at the users, but also enhances the precoding at the BS.
- To reap these cache-enabled benefits, low-complexity linear precoders based on maximum-ratio transmission (MRT), zero-forcing (ZF), and regularized zero-forcing (RZF) precoding are proposed. Different from conventional linear precoding which only requires channel state information (CSI) at the BS, the proposed linear precoders depend on both the CSI and the cache status. We analyze the performance of cache-aided massive MIMO for each considered linear precoding scheme. Lower bounds on the ergodic achievable rates of MRT and ZF precoding are derived in closed form. Additionally, we analyze the asymptotic performance of RZF precoding based on random matrix theory.
- Simulation results show that, compared to *conventional* massive MIMO, the proposed cache-aided massive MIMO scheme achieves a significantly higher ergodic rate, especially when ρ_0 is small.

We note that for notational convenience, we only consider a simple caching policy where each user caches entire files. However, the proposed precoding techniques are applicable to more general caching policies and may achieve even higher throughputs if each user caches portions of each file as in [11], [12]. Due to the limited page space, the design of throughput-optimal caching policies for the proposed cache-aided massive MIMO scheme and the corresponding performance analysis are deferred to future work.

The remainder of the paper is organized as follows. In Section II, the system model and the proposed cache-aided massive MIMO scheme are presented. We analyze the achievable rates of the proposed scheme for different linear precoders in Section III. The performance of cache-aided massive MIMO is evaluated in Section IV, and finally, Section V concludes the paper.

Notations: In this paper, we use boldface capital and lower case letters to denote matrices and vectors, respectively. $\mathbf{A}^{\mathrm{H}}$ and $\mathbf{A}^{\mathrm{T}}$ represent the complex conjugate transpose and the transpose of matrix $\mathbf{A}$, respectively; $\mathbf{A}^{-1}$ is the inverse of square matrix $\mathbf{A}$; $\mathrm{Pr}(\cdot)$ and $\mathcal{E}\{\cdot\}$ are the probability and the expectation operators, respectively; $\Re(\cdot)$ and $\Im(\cdot)$ represent the real and imaginary parts of a complex number, respectively. $\| \cdot \|$ and $| \cdot |$ are the Euclidean norm of a vector and the absolute value of a scalar, respectively. C_n^k is the k-out-of-n binomial coefficient. $A \to B$ indicates that A converges to B in the limit. $\mathcal{CN}\left(0, \sigma^2\right)$ denotes the complex Gaussian distribution with zero mean and variance σ^2, and finally, $\mathbb{C}^{N_r \times N_t}$ is the set of complex-valued $N_r \times N_t$ matrices.
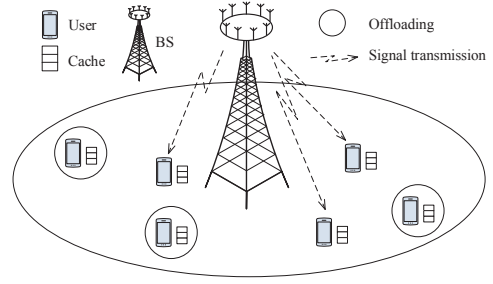

Fig. 1. Delivery model for cache-aided massive MIMO system, where each user obtains the requested file from its cache or the BS, depending on its cache status.

II. CACHE-AIDED MASSIVE MIMO

In this section, the system model, the interference cancellation mechanism, and the ergodic achievable rate of the proposed cache-aided massive MIMO scheme are presented.

A. System Model

As shown in Fig. 1, we consider a single-cell downlink system with an M-antenna BS and K single-antenna users. The BS stores a library of L_b popular files, where each file has a size of F MBytes. Each user is equipped with a cache memory of size $L_u F$ MBytes, where $L_u \leq L_b$, i.e., the cache capacity of each user is insufficient to store the whole library and only a portion of the files can be cached.

The system operates in two phases: a placement phase and a delivery phase. In the placement phase, all the users place L_u arbitrary files from the library into their own cache prior to the time of request. This phase may happen during the early mornings when cellular traffic is low. In the delivery phase, each user may request one of the L_b files. Let $c_{k,l} = 0$ if the file requested by user k has been cached at user l and $c_{k,l} = 1$ otherwise. Using this notation, the number of active transmission users in the proposed system is given by $\overline{K} = \sum_{k=1}^{K} c_{k,k}$. If user k has cached the file it requests, i.e., $c_{k,k} = 0$, it is fetched from its cache instantly. In this case, user k is considered inactive as it requires no cellular transmission. Otherwise, if $c_{k,k} = 1$, the requested file has to be transmitted by the BS, and user k is considered to be active.

B. Cache-Enabled Interference Cancellation

In this paper, we assume that the cache status is given and we focus on exploiting the cached data to improve the delivery to all the users. Assume that the file requested by user k is not cached at user k, i.e., $c_{k,k}=1$ and the BS has to transmit the file. Then, the received signal at user k, denoted by y_k, is given by

$$y_k = \mathbf{h}_k^{\mathrm{H}} \mathbf{w}_k s_k + \sum_{l \neq k} c_{l,l} \mathbf{h}_k^{\mathrm{H}} \mathbf{w}_l s_l + v_k, \qquad (1)$$

where s_k is the transmit symbol intended for user k with $\mathcal{E}\{|s_k|^2\} = E_k$. $\mathbf{w}_k \in \mathbb{C}^{M \times 1}$ is the precoding vector of user k and v_k is the additive white Gaussian noise following distribution $\mathcal{CN}\left(0, \sigma^2\right)$. $\mathbf{h}_k = [h_{k,1}, h_{k,2}, \cdots, h_{k,M}]^{\mathrm{T}} \in \mathbb{C}^{M \times 1}$ is the channel vector from the BS to user k. In this paper, channel coefficient $h_{k,m}$ is modeled as

$$h_{k,m} = g_{k,m}\sqrt{\beta_k}, \qquad (2)$$

where $g_{k,m}$ is the fading coefficient from the mth BS antenna to user k and follows distribution $\mathcal{CN}(0,1)$. β_k models the pathloss and shadowing effects and remains constant over a large number of coherence time intervals. We assume that the total transmit power at the BS is E_0. To satisfy the total transmit power constraint, we let $\sum_{k=1}^{K} c_{k,k} E_k = E_0$ and $\|\mathbf{w}_k\|^2 = 1$, $\forall k$.

If $c_{l,l} = 0$, user l is inactive and hence, it is offloaded for transmission in (1). On the other hand, if $c_{l,l} = 1$ and $c_{l,k} = 0$, i.e., the file requested by active user l is not cached at user l but is cached at user k, this cached file can still be exploited for interference cancellation [10]. In particular, by re-encoding this cached file and subtracting the corresponding signal from y_k, the interference caused by user l to user k can be removed[1] at user k. Consequently, by caching at the user side, user l with $c_{l,l} = 1$ causes interference only to users k with $c_{l,k} = 1$, $l \neq k$. Let $U_k \triangleq \{l \mid c_{l,l} = 1, c_{l,k} = 1, l \neq k\}$ be the set of users interfering with user k. The cardinality of U_k is denoted by N_k. The residual received signal of user k after interference cancellation, denoted by y_k^{IC}, is given by

$$y_k^{\text{IC}} = \mathbf{h}_k^{\text{H}} \mathbf{w}_k s_k + \sum_{l \in U_k} \mathbf{h}_k^{\text{H}} \mathbf{w}_l s_l + v_k. \tag{3}$$

C. Ergodic Achievable Rate

Based on (3), the SINR of user k is given by

$$\text{SINR}_k = \frac{\left|\mathbf{h}_k^{\text{H}} \mathbf{w}_k\right|^2 E_k}{\sum_{l \in U_k} \left|\mathbf{h}_k^{\text{H}} \mathbf{w}_l\right|^2 E_l + \sigma^2}. \tag{4}$$

Assume that the file delivery for each active user spans a large number of coherence time intervals. Then, with the proposed cache-aided massive MIMO, the ergodic achievable rate of user k is [2]

$$R_k = \mathcal{E}\left\{\log_2\left(1 + \text{SINR}_k\right)\right\}$$
$$= \mathcal{E}\left\{\log_2\left(1 + \frac{\left|\mathbf{h}_k^{\text{H}} \mathbf{w}_k\right|^2 E_k}{\sum_{l \in U_k} \left|\mathbf{h}_k^{\text{H}} \mathbf{w}_l\right|^2 E_l + \sigma^2}\right)\right\}. \tag{5}$$

Since $f(x) = \log_2(1 + \frac{1}{x})$ is a convex function, by employing Jensen's inequality, a lower bound on the ergodic achievable rate R_k is obtained as [2]

$$R_k \geq \tilde{R}_k \triangleq \log_2\left(1 + \left(\mathcal{E}\left\{\frac{\sum_{l \in U_k} \left|\mathbf{h}_k^{\text{H}} \mathbf{w}_l\right|^2 E_l + \sigma^2}{\left|\mathbf{h}_k^{\text{H}} \mathbf{w}_k\right|^2 E_k}\right\}\right)^{-1}\right). \tag{6}$$

III. CACHE-AIDED LINEAR PRECODER DESIGN

In this section, we investigate advanced precoder designs at the BS exploiting user-side caching for enhanced performance. We consider linear precoding techniques, namely MRT, ZF, and RZF, which are preferred in practical massive MIMO systems as they attain high performance with affordable computational complexity. We analyze the ergodic achievable rates for MRT and ZF precoding and derive corresponding lower bounds in

[1]For suppressing the interference caused by user l at user k, the re-encoded signal s_l is scaled by $\mathbf{h}_k^{\text{H}} \mathbf{w}_l$ before being subtracted from the received signal. Here, $\mathbf{h}_k^{\text{H}} \mathbf{w}_l$ can be estimated locally at user k requiring no knowledge about the requests and cache status of the other users.

Sections III-A and III-B, respectively. In Section III-C, we analyze the achievable rate for RZF precoding in the asymptotic regime, where $M, K \to \infty$ but ρ_0 is fixed. We assume that the CSI, $\mathbf{h}_k$, is perfectly known at both the receivers and the transmitter. Here, we consider $M > K$, while the proposed scheme is also applicable for $M \leq K$.

A. Maximum-Ratio Transmission

MRT precoding ensures that the signals transmitted by the BS over different antennas add up constructively at the intended user, and hence, maximizes the received signal power. However, when ρ_0 is small, MRT suffers from severe multiuser interference. In this case, cache-aided interference cancellation and offloading can improve the performance of MRT. With MRT and perfect CSI, the precoding vector of user k is [4]

$$\mathbf{w}_k^{\text{MRT}} = \frac{\mathbf{h}_k}{\|\mathbf{h}_k\|}. \tag{7}$$

By substituting (7) into (5), the ergodic achievable rate of user k with MRT precoding is

$$R_k^{\text{MRT}} = \mathcal{E}\left\{\log_2\left(1 + \frac{\|\mathbf{h}_k\|^2 E_k}{\sum_{l \in U_k} \frac{\left|\mathbf{h}_k^{\text{H}} \mathbf{h}_l\right|^2}{\|\mathbf{h}_l\|^2} E_l + \sigma^2}\right)\right\}. \tag{8}$$

Proposition 1. With MRT precoding in (7) and perfect CSI, the ergodic achievable rate of user k is lower bounded as

$$R_k^{\text{MRT}} \geq \tilde{R}_k^{\text{MRT}} = \log_2\left(1 + \frac{\beta_k(M-1)E_k}{\sum_{l \in U_k} \beta_k E_l + \sigma^2}\right). \tag{9}$$

Proof: Please refer to Appendix A. ∎

In Proposition 1, fading is averaged out in $\tilde{R}_k^{\text{MRT}}$. Moreover, as the desired and the interfering signals experience the same downlink channel, the lower bound on the ergodic achievable rate of a user depends only on the pathloss and shadowing of its own channel. The sum ergodic achievable rate of all users can be improved by optimizing the power allocation given the pathloss and shadowing of the different users. However, to illustrate the performance gains enabled by caching, we simply assume uniform transmit power allocation for all active users, where $E_k = E_0/\overline{K}$. Consequently, the lower bound in (9) simplifies to

$$\tilde{R}_k^{\text{MRT,uni}} = \log_2\left(1 + \frac{\beta_k(M-1)E_0}{\beta_k N_k E_0 + \overline{K}\sigma^2}\right). \tag{10}$$

Based on (10), the ergodic achievable rate of *conventional* massive MIMO (i.e., $N_k = K-1$, $\overline{K} = K$) for uniform transmit power allocation is lower bounded by

$$\tilde{R}_k^{\text{b,MRT,uni}} = \log_2\left(1 + \frac{\beta_k(M-1)E_0}{\beta_k(K-1)E_0 + K\sigma^2}\right). \tag{11}$$

Remark 1. In (10), N_k and $\overline{K}$ are proportional to K. This implies that both $\tilde{R}_k^{\text{MRT,uni}}$ and $\tilde{R}_k^{\text{b,MRT,uni}}$ increase monotonically with the number of BS antennas per user. Moreover, we have $\tilde{R}_k^{\text{MRT,uni}} \geq \tilde{R}_k^{\text{b,MRT,uni}}$, i.e., (11) defines the worst-case performance of the proposed scheme. Comparing (10) and (11), the performance gains of the proposed scheme over *conventional* massive MIMO include: i) an enhanced transmit

power for the active users due to cache-enabled offloading, when $\overline{K} < K$, and ii) reduced interference due to cache-enabled interference cancellation and offloading, when $N_k < K - 1$.

B. Zero-Forcing Precoding

Different from MRT, ZF precoding avoids multiuser interference by projecting the transmit signal of each user into the null space of all other users. When files are cached at the users' terminals, the interference cancellation capabilities offered by caching and ZF precoding can be combined for improved precoding design. In particular, if $c_{l,l} = 0$, i.e., user $l \neq k$ is inactive, its channel will not be considered for the ZF precoding design at user k; on the other hand, if user l is active and has cached the file requested by user k, i.e., $c_{l,l} = 1$ and $c_{k,l} = 0$, user l can exploit the cached file to remove the interference caused by user k, without having to rely on ZF precoding. Hence, the ZF percoder intended for user k, only needs to avoid causing interference to the set of active users that do not have user k's requested file in their caches. This set of users is denoted by $\Lambda_k \triangleq \{l \mid c_{l,l} = 1, c_{k,l} = 1, l \neq k\}$. Consequently, if Λ_k is not empty, the precoding vector $\mathbf{w}_k^{\mathrm{ZF}}$ of user k has to satisfy the following constraints:

$$\left\| \mathbf{w}_k^{\mathrm{ZF}} \right\|^2 = 1, \text{ and } \mathbf{h}_l^{\mathrm{H}} \mathbf{w}_k^{\mathrm{ZF}} = 0, l \in \Lambda_k, \qquad (12)$$

such that the signal of user k is sent in the null space of the signal space formed by the users in Λ_k. Let D_k and $\Lambda_k(n)$ be the cardinality and the nth element of set Λ_k, respectively. Then, for user k and set Λ_k, we define the effective channel matrix after cache-enabled interference cancellation as

$$\mathbf{Q}_k = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{D_k+1}], \qquad (13)$$

where $\mathbf{q}_1 = \mathbf{h}_k$, and $\mathbf{q}_{n+1} = \mathbf{h}_{\Lambda_k(n)}$, $n = 1, 2, \cdots, D_k$. Consequently, for the proposed cache-aided massive MIMO, the ZF precoding vector of user k is given by

$$\mathbf{w}_k^{\mathrm{ZF}} = \frac{\mathbf{Q}_k(\mathbf{Q}_k^{\mathrm{H}} \mathbf{Q}_k)^{-1} \mathbf{e}_1}{\left\| \mathbf{Q}_k(\mathbf{Q}_k^{\mathrm{H}} \mathbf{Q}_k)^{-1} \mathbf{e}_1 \right\|}, \qquad (14)$$

where $\mathbf{e}_1 \triangleq [1, 0, \cdots, 0]^{\mathrm{T}}$. For the ZF precoder $\mathbf{w}_k^{\mathrm{ZF}}$, the corresponding ergodic achievable rate and its lower bound are given in Proposition 2.

Proposition 2. With the ZF precoder $\mathbf{w}_k^{\mathrm{ZF}}$ in (14) and perfect CSI, the ergodic achievable rate of user k is

$$R_k^{\mathrm{ZF}} = \mathcal{E} \left\{ \log_2 \left(1 + \frac{E_k}{\left\| \mathbf{Q}_k(\mathbf{Q}_k^{\mathrm{H}} \mathbf{Q}_k)^{-1} \mathbf{e}_1 \right\|^2 \sigma^2} \right) \right\}. \qquad (15)$$

Moreover, R_k^{ZF} is lower bounded by

$$R_k^{\mathrm{ZF}} \geq \tilde{R}_k^{\mathrm{ZF}} = \log_2 \left(1 + \frac{\beta_k (M - D_k - 1) E_k}{\sigma^2} \right). \qquad (16)$$

Proof: Please refer to Appendix B. ∎

If uniform transmit power allocation is adopted, the lower bound in (16) simplifies to

$$\tilde{R}_k^{\mathrm{ZF,uni}} = \log_2 \left(1 + \frac{\beta_k (M - D_k - 1) E_0}{\overline{K} \sigma^2} \right). \qquad (17)$$

Based on (17), the ergodic achievable rate of user k for *conventional* massive MIMO (i.e., $D_k = K - 1$, $\overline{K} = K$) is

lower bounded by

$$\tilde{R}_k^{\mathrm{b,ZF,uni}} = \log_2 \left(1 + \frac{\beta_k (M - K) E_0}{K \sigma^2} \right). \qquad (18)$$

Remark 2. We have $\tilde{R}_k^{\mathrm{ZF,uni}} \geq \tilde{R}_k^{\mathrm{b,ZF,uni}}$. To explain the performance difference, we note that, in the conventional ZF-based massive MIMO system, the signal of user k has to be orthogonal to the signals of all other $K-1$ users. The resulting interference mitigation comes at the cost of a reduced received signal power for each user. In contrast, with the proposed scheme, as cached-enabled offloading and interference cancellation can partially mitigate the interference, more spatial degrees of freedom are available for ZF precoding design and hence, the power loss incurred by ZF precoding is reduced. Moreover, due to cache-enabled offloading, a power gain of $K/\overline{K}$ is also achieved for transmit power allocation to the active users.

C. Regularized Zero-Forcing Precoding

In *conventional* massive MIMO, RZF precoding is often considered to balance between interference mitigation and power enhancement [16]. For the proposed scheme, RZF precoding has to be reconsidered in order to maximize the performance gains enabled by caching. However, the ergodic rate of RZF precoding cannot be analyzed in the same manner as that of MRT/ZF precoding. To make the analysis tractable, we investigate RZF precoding in the large system limit, when $M, K \to \infty$ but ρ_0 is fixed and we assume that the β_ks are equal[2].

We define the effective channel fading matrix for user k as

$$\mathbf{F}_k = [\mathbf{f}_{k,1}, \mathbf{f}_{k,2}, \cdots, \mathbf{f}_{k,D_k+1}]^{\mathrm{T}}, \qquad (19)$$

where $\mathbf{f}_{k,1} = \mathbf{g}_k$, $\mathbf{f}_{k,n+1} = \mathbf{g}_{\Lambda_k(n)}, n = 1, 2, \cdots, D_k$, and $\mathbf{g}_k = [g_{k,1}, g_{k,2}, \cdots, g_{k,M}]^{\mathrm{T}}$. Then, the RZF precoding vector of user k is given by [15], [16]

$$\mathbf{w}_k^{\mathrm{RZF}} = \frac{(\mathbf{F}_k^{\mathrm{H}} \mathbf{F}_k + \alpha_k \mathbf{I})^{-1} \mathbf{f}_{k,1}}{\left\| (\mathbf{F}_k^{\mathrm{H}} \mathbf{F}_k + \alpha_k \mathbf{I})^{-1} \mathbf{f}_{k,1} \right\|}, \qquad (20)$$

where α_k is a regularization parameter.

Proposition 3. With RZF precoding in (20) and perfect CSI, the ergodic achievable rate of user k is given by

$$R_k^{\mathrm{RZF}} = \log_2 \left(1 + \frac{E^{\mathrm{s}}}{\sum_{l \in U_k} E^{\mathrm{i}}(l) + \sigma^2} \right), \qquad (21)$$

where $E^{\mathrm{s}} \to \frac{-M \beta_k \mathcal{G}^2(\rho_k, \xi_k) E_k}{\frac{d}{d\xi_k} \mathcal{G}(\rho_k, \xi_k)}$ is the received signal power of user k, $E^{\mathrm{i}}(l) \to \frac{\beta_k E_l}{(1 + \mathcal{G}(\rho_l, \xi_l))^2}$ is the interference power received at user k and caused by user l. Moreover, $\rho_k = D_k/M$, $\xi_k = \alpha_k/M$, and $\mathcal{G}(\rho_k, \xi_k)$ can be evaluated in closed form [15]

$$\mathcal{G}(\rho_k, \xi_k) = \frac{1}{2} \left[\sqrt{\frac{(1 - \rho_k)^2}{\xi_k^2} + \frac{2(1 + \rho_k)}{\xi_k} + 1} + \frac{1 - \rho_k}{\xi_k} - 1 \right]. \qquad (22)$$

Proof: Please refer to Appendix C. ∎

---

[2]This assumption facilitates concise and insightful results. Nevertheless, the extension to non-identical β_ks is possible [16] and will be provided in the journal version of the paper.

If uniform transmit power allocation is adopted, the ergodic achievable rate of user k in (21) reduces to

$$R_k^{\text{RZF,uni}} = \log_2\left(1 + \frac{E^{\text{s,uni}}}{\sum\limits_{l \in U_k} E^{\text{i,uni}}(l) + \sigma^2}\right), \quad (23)$$

where $E^{\text{s,uni}} \rightarrow \frac{-M\beta_k \mathcal{G}^2(\rho_k,\xi_k)E_0}{\frac{d}{d\xi_k}\mathcal{G}(\rho_k,\xi_k)\overline{K}}$, $E^{\text{i,uni}}(l) \rightarrow \frac{\beta_k E_0}{(1+\mathcal{G}(\rho_l,\xi_l))^2\overline{K}}$. Similarly, the ergodic achievable rate of user k in a *conventional* massive MIMO system (i.e., $\rho_k = (K-1)/M$, $\xi_k = \alpha/M$, $\overline{K} = K$) can be obtained from (23), and is given by

$$R_k^{\text{b,RZF,uni}} = \log_2\left(1 + \frac{E^{\text{b,s,uni}}}{\sum\limits_{l \neq k} E^{\text{b,i,uni}}(l) + \sigma^2}\right), \quad (24)$$

where $E^{\text{b,s,uni}} \rightarrow \frac{-M\beta_k \mathcal{G}^2(\frac{K-1}{M},\frac{\alpha}{M})E_0}{\frac{d}{d\frac{\alpha}{M}}\mathcal{G}(\frac{K-1}{M},\frac{\alpha}{M})K}$, $E^{\text{b,i,uni}}(l) \rightarrow \frac{\beta_k E_0}{(1+\mathcal{G}(\frac{K-1}{M},\frac{\alpha}{M}))^2 K}$, and α is the regularization parameter in *conventional* massive MIMO.

Remark 3. Comparing $R_k^{\text{RZF,uni}}$ with $R_k^{\text{b,RZF,uni}}$, we observe that, by employing the proposed scheme, caching not only improves the transmit power by reducing the number of active users $\overline{K}$, but also impacts the signal and interference powers. Therefore, for RZF precoding, the tradeoff between the signal and the interference powers, which is adjusted by the regularization parameter, has to be newly investigated for maximization of the cache-enabled performance gains. In this paper, the optimal regularization parameter is found numerically for the results shown in Section IV.

IV. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed scheme. For comparison, *conventional* massive MIMO is adopted as a baseline. Let $\text{SNR} \triangleq 10\log_{10}\frac{E_0}{\sigma^2}$ be the transmit signal-to-noise ratio. We set $\text{SNR} = 10$ dB, $\beta_k = 0.5$, $\forall k$, $F = 1$ MByte, and $L_b = 100$. We assume that the files are requested with equal probability $1/L_b$. To illustrate the benefits of caching, we consider a simple uniform caching scheme, whereby each user caches a file with probability $p \triangleq L_u/L_b$. We note that, by considering random requests and caching, $\overline{K}$, N_k (i.e., the number of interfering users for user k), and D_k (i.e., the number of active users that do not have user k's requested file in their caches) become random variables. However, for performance evaluation, we consider the case of asymptotically large K. For large K, due to the law of large numbers, $\overline{K}$, N_k, and D_k converge to their mean values $\mathcal{E}\{\overline{K}\}$, $\mathcal{E}\{N_k\}$, and $\mathcal{E}\{D_k\}$, respectively. Hence, we have $\mathcal{E}\{\overline{K}\} = (1-p)K$. Furthermore, for a given user k, the event that user $l \neq k$ causes interference to user k, i.e., $l \in U_k$, has probability

$$p_u = \Pr(c_{l,l}=1, c_{k,k}=1, c_{l,k}=1, l \neq k) = p_u^{(1)} + p_u^{(2)}, \quad (25)$$

where $p_u^{(1)}$ is the probability that users l and k request different files, which is given by $p_u^{(1)} = C_{L_b}^1 C_{L_b-1}^1 \left[\frac{1}{L_b}(1-p)\right]^2(1-p)$. $p_u^{(2)}$ is the probability that the two users require the same file, given by $p_u^{(2)} = C_{L_b}^1 \left[\frac{1}{L_b}(1-p)\right]^2$. Consequently, $p_u = \left(1 - \frac{1}{L_b}\right)(1-p)^3 + \frac{1}{L_b}(1-p)^2$. Since p_u is independent of the users, we have $\mathcal{E}\{N_k\} = (K-1)p_u$. Following the same approach as
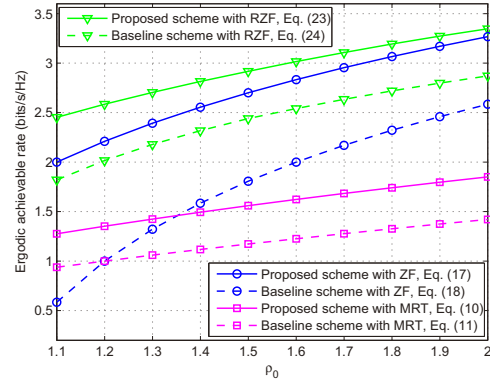

Fig. 2. Ergodic achievable rate per user vs. number of BS antennas per user, ρ_0, for MRT, ZF, and RZF precoders with $L_u = 20$, i.e., 20% of the users' requests are offloaded by caching.
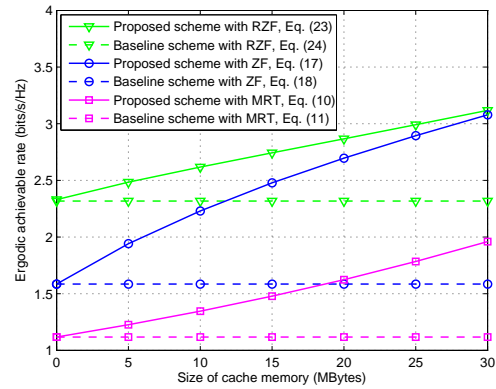

Fig. 3. Ergodic achievable rate per user vs. cache size, $L_u F$, for MRT, ZF, and RZF precoders with $\rho_0 = 1.4$.

for $\mathcal{E}\{N_k\}$, we can further show $\mathcal{E}\{D_k\} = (K-1)p_u$. To show the maximum performance of RZF precoding, the regularization parameter is optimized numerically for each parameter setting.

Fig. 2 illustrates the ergodic achievable rate[3] per user versus the number of BS antennas per user, ρ_0, for MRT, ZF precoding, and RZF precoding. From Fig. 2 we observe that, for all considered precoders, the proposed scheme achieves significantly higher ergodic rates than the baseline scheme. This is because, on the one hand, caching offloads the cellular traffic for inactive users and mitigates the multiuser interference of active users. On the other hand, with the enhanced precoders, caching is further exploited to improve the received signal power and/or increase the spatial degrees of freedom, both of which increase the ergodic achievable rate. For example, with MRT precoding, the proposed scheme for $\rho_0 = 1.1$ achieves the same performance as the baseline scheme for $\rho_0 = 1.8$. Moreover, by optimally balancing between interference cancellation and power enhancement, enabled by user-side caching and BS-side precoding, respectively, the proposed scheme with RZF precoding achieves the best performance.

From Fig. 2 we also observe that the proposed scheme achieves the largest performance gains over the baseline scheme for ZF precoding when the number of antennas approaches the

[3]The numerical results shown in this section were obtained with (10), (11), (17), (18), (23), and (24) for $M, K \rightarrow \infty$, $\rho_0 = M/K$, and have been validated by Monte Carlo simulations. However, for clarity, the simulation results are not included in Figs. 2 and 3.

number of users. This is because the cache-enabled interference cancellation is exploited at the BS to reduce the number of ZF precoding constraints. For a small ρ_0, the signal space for ZF precoding design is severely constrained. In this case, ZF precoding can benefit from the increased spatial degrees of freedom enabled by caching and hence achieve a large performance improvement. Thus, caching can effectively enhance the performance of massive MIMO systems having a small number of BS antennas per user, i.e., when ρ_0 is small.

Fig. 3 shows the ergodic achievable rate per user versus the cache size, $L_u F$, for MRT, ZF, and RZF precoding. We observe that, for all considered precoders, as the cache size increases, the proposed scheme can exploit the increased offloading and interference cancellation opportunities enabled by caching to significantly improve the system performance. For example, when $L_u = 20$, we have $\mathcal{E}\{\overline{K}\}/K = 0.8$, i.e., 20% of the users' requests are offloaded by caching, and $p_u = \mathcal{E}\{N_k\}/(K-1) = \mathcal{E}\{D_k\}/(K-1) = 0.513$, i.e., 28.7% of the users have cached the files requested by other users. Consequently, due to caching, 48.7% of the users including the offloaded users would not cause interference to the active users. In this case, for ZF precoding, the ergodic achievable rate of the proposed scheme increases by 70.1% compared to the baseline scheme. On the other hand, as caching is unavailable for the baseline scheme, its performance is independent of L_u. From Fig. 3 we also observe that, for the proposed scheme, RZF precoding achieves the best performance among the considered precoding techniques for all considered cache sizes.

V. CONCLUSION

In this paper, a novel cache-aided massive MIMO scheme was proposed. In addition to reaping the advantages of caching and massive MIMO, the proposed scheme also facilitates interference cancellation at the user side and transmit power savings at the BS. Exploiting these cache-enabled benefits, linear precoders, specifically MRT, ZF, and RZF precoders, were redesigned for further performance improvement. Closed-form expressions for the ergodic achievable rate of the proposed schemes were derived for MRT, ZF, and RZF precoding. Numerical results show that the proposed scheme significantly improves the performance of all considered precoding techniques especially when the number of BS antennas per user is small.

APPENDIX

A. Proof of Proposition 1

Substituting (7) into (6), we have

$$\tilde{R}_k = \log_2\left(1 + \left(\mathcal{E}\left\{\frac{\sum_{l \in U_k} \frac{|\mathbf{h}_k^H \mathbf{h}_l|^2}{\|\mathbf{h}_l\|^2} E_l + \sigma^2}{\|\mathbf{h}_k\|^2 E_k}\right\}\right)^{-1}\right), \quad (26)$$

where [2, Appendix A],

$$\mathcal{E}\left\{\frac{\sum_{l \in U_k} \frac{|\mathbf{h}_k^H \mathbf{h}_l|^2}{\|\mathbf{h}_l\|^2} E_l + \sigma^2}{\|\mathbf{h}_k\|^2 E_k}\right\} = \left(\sum_{l \in U_k} \beta_k E_l + \sigma^2\right) \mathcal{E}\left\{\frac{1}{\|\mathbf{h}_k\|^2 E_k}\right\}. \quad (27)$$

Due to (2), we have $\|\mathbf{h}_k\|^2 = \beta_k \sum_{m=1}^M |g_{k,m}|^2$. Note that, $|g_{k,m}|^2 = |\Re(g_{k,m})|^2 + |\Im(g_{k,m})|^2$. Hence, $2\sum_{m=1}^M |g_{k,m}|^2$ follows the chi-squared distribution with $2M$ degrees of freedom as $\sqrt{2}\Re(g_{k,m})$ and $\sqrt{2}\Im(g_{k,m})$ are independent standard normal random variables. Thus, $2\sum_{m=1}^M |g_{k,m}|^2$ is an inverse chi-square distribution with $2M$ degrees of freedom, and we have $\mathcal{E}\left\{\left(\sum_{m=1}^M |g_{k,m}|^2\right)^{-1}\right\} = (M-1)^{-1}$, and $\mathcal{E}\left\{1/\|\mathbf{h}_k\|^2\right\} = \frac{1}{\beta_k(M-1)}$. Substituting (27) into (26), Proposition 1 is proved.

B. Proof of Proposition 2

Substituting (14) into (5), we have

$$R_k^{ZF} = \mathcal{E}\left\{\log_2\left(1 + \frac{E_k}{\sigma^2} \frac{|\mathbf{h}_k^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1|^2}{\|\mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1\|^2}\right)\right\}. \quad (28)$$

Since $|\mathbf{q}_1^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_l|^2 = 0$, $l = 2, \cdots, D_k+1$, we have $|\mathbf{q}_1^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1|^2 = \|\mathbf{q}_1^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1}\|^2 = \mathbf{q}_1^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{Q}_k^H \mathbf{q}_1$. Note that $\mathbf{q}_1^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{Q}_k^H \mathbf{q}_1$ is the element in the first row and the first column of $\mathbf{Q}_k^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{Q}_k^H \mathbf{Q}_k = \mathbf{I}$. Hence, we have $|\mathbf{h}_k^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1|^2 = 1$ as $\mathbf{h}_k = \mathbf{q}_1$. Then, substituting $|\mathbf{h}_k^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1|^2 = 1$ into (28), (15) in Proposition 2 is proved.

Moreover, substituting (15) into (6), the downlink achievable rate of user k is lower bounded by

$$\tilde{R}_k^{ZF} = \log_2\left(1 + \frac{E_k}{\sigma^2}\left(\mathcal{E}\left\{\|\mathbf{Q}_k(\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1\|^2\right\}\right)^{-1}\right). \quad (29)$$

Since $\|\mathbf{Q}_k(\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1\|^2 = \mathbf{e}_1^H (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{Q}_k^H \mathbf{Q}_k (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1 = \mathbf{e}_1^H (\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1$, we have that $\|\mathbf{Q}_k(\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1\|^2$ is the element in the first row and the first column of matrix $(\mathbf{Q}_k^H \mathbf{Q}_k)^{-1}$. Note that $(\mathbf{Q}_k^H \mathbf{Q}_k)^{-1}$ is a complex inverse Wishart matrix with D_k+1 degrees of freedom and parameter matrix $\mathbf{\Phi}^{-1}$ [13], where $\mathbf{\Phi} \in \mathbb{C}^{(D_k+1) \times (D_k+1)}$ is a diagonal matrix with diagonal elements $[\beta_k, \beta_{\Lambda_k(1)}, \cdots, \beta_{\Lambda_k(D_k)}]$. Using the results in [14, Ch. 3.8], we have $\mathcal{E}\left\{(\mathbf{Q}_k^H \mathbf{Q}_k)^{-1}\right\} = \frac{\mathbf{\Phi}^{-1}}{M-(D_k+1)}$. Thus, we have

$$\mathcal{E}\left\{\|\mathbf{Q}_k(\mathbf{Q}_k^H \mathbf{Q}_k)^{-1} \mathbf{e}_1\|^2\right\} = \frac{1}{(M - D_k - 1)\beta_k}. \quad (30)$$

Then, substituting (30) into (29), (16) in Proposition 2 is proved.

C. Proof of Proposition 3

Based on (4) and (20), the effective signal power is given as

$$E^s = |\mathbf{h}_k^H \mathbf{w}_k^{RZF}|^2 E_k = \frac{|\mathbf{h}_k^H (\mathbf{F}_k^H \mathbf{F}_k + \alpha_k \mathbf{I})^{-1} \mathbf{f}_{k1}|^2 E_k}{\|(\mathbf{F}_k^H \mathbf{F}_k + \alpha_k \mathbf{I})^{-1} \mathbf{f}_{k1}\|^2}$$

$$= \frac{\left| \mathbf{g}_k^{\mathrm{H}} (\mathbf{F}_k^{\mathrm{H}} \mathbf{F}_k + \alpha_k \mathbf{I})^{-1} \mathbf{g}_k \right|^2 \beta_k E_k}{\left\| (\mathbf{F}_k^{\mathrm{H}} \mathbf{F}_k + \alpha_k \mathbf{I})^{-1} \mathbf{g}_k \right\|^2}. \tag{31}$$

Applying the matrix inversion lemma [15], we have

$$(\mathbf{F}_k^{\mathrm{H}} \mathbf{F}_k + \alpha_k \mathbf{I})^{-1} \mathbf{g}_k = \frac{(\mathbf{F}_{k(k)}^{\mathrm{H}} \mathbf{F}_{k(k)} + \alpha_k \mathbf{I})^{-1} \mathbf{g}_k}{1 + \mathbf{g}_k^{\mathrm{H}} (\mathbf{F}_{k(k)}^{\mathrm{H}} \mathbf{F}_{k(k)} + \alpha_k \mathbf{I})^{-1} \mathbf{g}_k}, \tag{32}$$

where $\mathbf{F}_{k(k)}$ is obtained by deleting the vector $\mathbf{g}_k$ from $\mathbf{F}_k$. Defining

$$X_k = \mathbf{g}_k^{\mathrm{H}} (\mathbf{F}_{k(k)}^{\mathrm{H}} \mathbf{F}_{k(k)} + \alpha_k \mathbf{I})^{-1} \mathbf{g}_k, \tag{33}$$

$$\boldsymbol{\Phi}_k = (\mathbf{F}_{k(k)}^{\mathrm{H}} \mathbf{F}_{k(k)} + \alpha_k \mathbf{I})^{-1}, \tag{34}$$

with (32), we have

$$E^{\mathrm{s}} = \frac{| X_k |^2 \beta_k E_k}{\left\| \boldsymbol{\Phi}_k \mathbf{g}_k \right\|^2}. \tag{35}$$

Rewrite X_k as $X_k = \frac{1}{M} \mathbf{g}_k^{\mathrm{H}} (\frac{1}{M} \mathbf{F}_{k(k)}^{\mathrm{H}} \mathbf{F}_{k(k)} + \xi_k \mathbf{I})^{-1} \mathbf{g}_k$, where $\xi_k = \alpha_k / M$. Then, in the large system limit where $M, D_k \to \infty$, but $\rho_k = D_k / M$ is finite and fixed, X_k converges (almost surely) to [15], [17]

$$\mathcal{G}(\rho_k, \xi_k) = \int_0^\infty \frac{1}{\mu + \xi_k} d\mathcal{F}_{\rho_k}(\mu), \tag{36}$$

where

$$\mathcal{F}_{\rho_k}(\mu) \triangleq (1 - \rho_k)^+ \delta(\mu) + \frac{\sqrt{(\mu - (1 - \sqrt{\rho_k})^2)^+ ((1 + \sqrt{\rho_k})^2 - \mu)^+}}{2\pi\mu}, \tag{37}$$

$\delta(\mu)$ is the impulse function, and $(x)^+ \triangleq \max\{0, x\}$. On the other hand, define $\bar{\boldsymbol{\Phi}}_k = (\frac{1}{M} \mathbf{F}_{k(k)}^{\mathrm{H}} \mathbf{F}_{k(k)} + \xi_k \mathbf{I})^{-1}$. Then, we have

$$\left\| \boldsymbol{\Phi}_k \mathbf{g}_k \right\|^2 = \left| \mathbf{g}_k^{\mathrm{H}} \boldsymbol{\Phi}_k^{\mathrm{H}} \boldsymbol{\Phi}_k \mathbf{g}_k \right| = \frac{1}{M} \left| \frac{1}{M} \mathbf{g}_k^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_k \bar{\boldsymbol{\Phi}}_k \mathbf{g}_k \right|. \tag{38}$$

Following a similar approach as for X_k, one can show that

$$\frac{1}{M} \mathbf{g}_k^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_k \bar{\boldsymbol{\Phi}}_k \mathbf{g}_k \to \int_0^\infty \frac{1}{(\mu + \xi_k)^2} \mathcal{F}_{\rho_k}(\mu) d\mu \tag{39}$$

$$\to -\frac{d}{d\xi_k} \mathcal{G}(\rho_k, \xi_k), \tag{40}$$

where the second line holds since $\frac{1}{(\mu + \xi_k)^2} = -\frac{d}{d\xi_k} \frac{1}{(\mu + \xi_k)}$. Substituting (36) and (40) into (35), we get

$$E^{\mathrm{s}} \to \frac{\mathcal{G}^2(\rho_k, \xi_k) \beta_k E_k}{-\frac{1}{M} \frac{d}{d\xi_k} \mathcal{G}(\rho_k, \xi_k)} = \frac{-M \mathcal{G}^2(\rho_k, \xi_k) \beta_k E_k}{\frac{d}{d\xi_k} \mathcal{G}(\rho_k, \xi_k)}. \tag{41}$$

If $l \in U_k$, the effective interference power of user l to user k is given by

$$E^{\mathrm{i}}(l) = |\mathbf{h}_k^{\mathrm{H}} \mathbf{w}_l^{\mathrm{RZF}}|^2 E_l = \frac{|\mathbf{h}_k^{\mathrm{H}} (\mathbf{F}_l^{\mathrm{H}} \mathbf{F}_l + \alpha_l \mathbf{I})^{-1} \mathbf{f}_{l,1}|^2 E_l}{\| (\mathbf{F}_l^{\mathrm{H}} \mathbf{F}_l + \alpha_l \mathbf{I})^{-1} \mathbf{f}_{l,1} \|^2}$$
$$= \frac{|\mathbf{g}_k^{\mathrm{H}} (\mathbf{F}_l^{\mathrm{H}} \mathbf{F}_l + \alpha_l \mathbf{I})^{-1} \mathbf{g}_l|^2 \beta_k E_l}{\| (\mathbf{F}_l^{\mathrm{H}} \mathbf{F}_l + \alpha_l \mathbf{I})^{-1} \mathbf{g}_l \|^2}. \tag{42}$$

Applying the matrix inversion lemma, (42) is rewritten as

$$E^{\mathrm{i}}(l) = \frac{|\mathbf{g}_k^{\mathrm{H}} \boldsymbol{\Phi}_l \mathbf{g}_l|^2 \beta_k E_l}{\| \boldsymbol{\Phi}_l \mathbf{g}_l \|^2}. \tag{43}$$

Removing $\mathbf{g}_k$ from $\mathbf{F}_{l(l)}$ and applying the matrix inversion lemma, we have

$$|\mathbf{g}_k^{\mathrm{H}} \boldsymbol{\Phi}_l \mathbf{g}_l|^2 = \frac{|\mathbf{g}_k^{\mathrm{H}} (\mathbf{F}_{l(lk)}^{\mathrm{H}} \mathbf{F}_{l(lk)} + \alpha_l \mathbf{I})^{-1} \mathbf{g}_l|^2}{|1 + \mathbf{g}_k^{\mathrm{H}} (\mathbf{F}_{l(lk)}^{\mathrm{H}} \mathbf{F}_{l(lk)} + \alpha_l \mathbf{I})^{-1} \mathbf{g}_k|^2}, \tag{44}$$

where $\mathbf{F}_{l(lk)}$ is obtained by deleting vectors $\mathbf{g}_k$ and $\mathbf{g}_l$ from $\mathbf{F}_l$. Therein, for the numerator, we have

$$|\mathbf{g}_k^{\mathrm{H}} \boldsymbol{\Phi}_{l(k)} \mathbf{g}_l|^2 = \frac{1}{M^2} \mathbf{g}_k^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_{l(k)} \mathbf{g}_l \mathbf{g}_l^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_{l(k)}^{\mathrm{H}} \mathbf{g}_k$$
$$\to \frac{1}{M} \mathrm{tr}(\frac{1}{M} \bar{\boldsymbol{\Phi}}_{l(k)} \mathbf{g}_l \mathbf{g}_l^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_{l(k)}^{\mathrm{H}}) = \frac{1}{M^2} \mathbf{g}_l^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_{l(k)}^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_{l(k)} \mathbf{g}_l$$
$$\to \frac{1}{M} \int_0^\infty \frac{1}{(\mu + \xi_l)^2} \mathcal{F}_{\rho_l^-}(\mu) d\mu \to -\frac{1}{M} \frac{d}{d\xi_l} \mathcal{G}(\rho_l^-, \xi_l). \tag{45}$$

where $\rho_l^- = (N_l - 1)/M$, $\boldsymbol{\Phi}_{l(k)} = (\mathbf{F}_{l(lk)}^{\mathrm{H}} \mathbf{F}_{l(lk)} + \alpha_l \mathbf{I})^{-1}$, and $\bar{\boldsymbol{\Phi}}_{l(k)} = (\frac{1}{M} \mathbf{F}_{l(lk)}^{\mathrm{H}} \mathbf{F}_{l(lk)} + \frac{\alpha_l}{M} \mathbf{I})^{-1}$. Moreover, for the denominator, based on (36), we have

$$|1 + \mathbf{g}_k^{\mathrm{H}} (\mathbf{F}_{l(lk)}^{\mathrm{H}} \mathbf{F}_{l(lk)} + \alpha_l \mathbf{I})^{-1} \mathbf{g}_k|^2 \to (1 + \mathcal{G}(\rho_l^-, \xi_l))^2. \tag{46}$$

As $M, N_l \to \infty$, we have $\rho_l^- \to \rho_l$, and hence, substituting (45) and (46) into (44), we get that

$$|\mathbf{g}_k^{\mathrm{H}} \boldsymbol{\Phi}_l \mathbf{g}_l|^2 \to \frac{-\frac{1}{M} \frac{d}{d\xi_l} \mathcal{G}(\rho_l, \xi_l)}{(1 + \mathcal{G}(\rho_l, \xi_l))^2}. \tag{47}$$

Based on (40), we have

$$\| \boldsymbol{\Phi}_l \mathbf{g}_l \|^2 = \frac{1}{M} \left(\frac{1}{M} \mathbf{g}_l^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_l^{\mathrm{H}} \bar{\boldsymbol{\Phi}}_l \mathbf{g}_l \right) \to -\frac{1}{M} \frac{d}{d\xi_l} \mathcal{G}(\rho_l, \xi_l), \tag{48}$$

which holds for $M \to \infty$. Hence, we have

$$E^{\mathrm{i}}(l) \to \frac{\beta_k E_l}{(1 + \mathcal{G}(\rho_l, \xi_l))^2}. \tag{49}$$

Moreover, the noise power is σ^2. Hence, when $D_k, M \to \infty$, but ρ_k is finite and fixed, the SINR of user k is given by

$$\mathrm{SINR}_k = \frac{E^{\mathrm{s}}}{\sum_{l \in U_k} E^{\mathrm{i}}(l) + \sigma^2}. \tag{50}$$

Substituting (41), (49), and (50) into (5), Proposition 3 is proved.

REFERENCES

[1] V. W. S. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*, Cambridge University Press, 2017.

[2] 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.

[3] H. Huh, S. Moon, Y. Kim, I. Lee, and G. Caire, "Multi-cell MIMO downlink with cell cooperation and fair scheduling: A large system limit analysis," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. 7771-7786, Dec. 2011.

[4] H. Yang and T. L. Marzetta, "Performance of conjugate and zeroforcing beamforming in large-scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 172-179, Feb. 2013.

[5] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Massive MU-MIMO downlink TDD systems with linear precoding and downlink pilots," *in Proc. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, Oct. 2013.

[6] O. Raeesi, A. Gokceoglu, Y. Zou, E. Björnson, and M. Valkama, "Performance analysis of multi-user massive MIMO downlink under channel non-reciprocity and imperfect CSI," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2456-2471, June 2018.

[7] W. Yu, "On the fundamental limits of massive connectivity," *in Proc. Inf. Theory App. (ITA) Workshop*, San Diego, CA, Feb. 2017.

[8] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114-123, Feb. 2016.

[9] O. Semiari, W. Saad, and M. Bennis, "Caching meets millimeter wave communications for enhanced mobility management in 5G networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 779-793, Feb. 2018.

[10] L. Xiang, D. W. K. Ng, X. Ge, Z. Ding, V. W. S. Wong, and R. Schober, "Cache-aided non-orthogonal multiple access," *in Proc. IEEE ICC*, Kansas City, USA, May 2018.

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

[12] 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.

[13] A. M. Tulino and S. Verdu, "Random matrix theory and wireless communications," *Foundations Trends Commun. Inf. Theory*, vol. 1, no. 1, pp. 1-182, June 2004.

[14] K. V. Mardia, J. T. Kent, and J. M. Bibby, "Multivariate analysis," *Academic Press.*, Academic Press, 1979.

[15] V. K. Nguyen and J. S. Evans, "Multiuser transmit beamforming via regularized channel inversion: A large system analysis," *in Proc. IEEE Global Commun. Conf.*, New Orleans, LA, USA, Dec. 2008.

[16] J. Zhu, R. Schober, and V. K. Bhargava, "Linear precoding of data and artificial noise in secure massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2245-2261, Mar. 2016.

[17] J. S. Evans and D. N. C. Tse, "Large system performance of linear multiuser receivers in multipath fading channels," *IEEE Trans. Inf. Theory*, vol. 46, pp. 2059-2078, Sep. 2000.