Mixed-Timescale Precoding and Cache Control in Cached MIMO Interference Network

An Liu, Member IEEE, and Vincent Lau, Fellow IEEE,
Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology

Abstract—Consider media streaming in MIMO interference networks whereby multiple base stations (BS) simultaneously deliver media to their associated users using fixed data rates. The performance is fundamentally limited by the cross-link interference. We propose a cache-induced opportunistic cooperative MIMO (CoMP) for interference mitigation. By caching a portion of the media files, the BSs opportunistically employ CoMP to transform the cross-link interference into spatial multiplexing gain. We study a mixed-timescale optimization of MIMO precoding and cache control to minimize the transmit power under the rate constraint. The cache control is to create more CoMP opportunities and is adaptive to the long-term popularity of the media files. The precoding is to guarantee the rate requirement and is adaptive to the channel state information and cache state at the BSs. The joint stochastic optimization problem is decomposed into a short-term precoding and a long-term cache control problem. We propose a precoding algorithm which converges to a stationary point of the short-term problem. Based on this, we exploit the hidden convexity of the long-term problem and propose a low complexity and robust solution using stochastic subgradient. The solution has significant gains over various baselines and does not require explicit knowledge of the media popularity.

Index Terms—Wireless media streaming, Dynamic cache control, Opportunistic CoMP, MIMO Precoding

I. INTRODUCTION

Media streaming is going to be one of the major applications in wireless networks. For example, it is envisioned that a significant portion of the capacity demand in future wireless systems will come from media streaming applications. In this paper, we consider media streaming in MIMO interference networks whereby multiple BSs simultaneously deliver media to their associated users using fixed data rates. The performance of this system is fundamentally limited by the inter-cell interference from the cross-links. In traditional cellular networks, the inter-cell interference is mitigated using frequency planning techniques such as frequency reuse or fractional frequency reuse [1]. To further improve the spectrum efficiency, more advanced techniques such as cooperative MIMO (CoMP) [2] and coordinated MIMO [3] have been proposed for future wireless systems. The CoMP technique can transform the cross-link interference into spatial multiplexing gain by sharing both real-time channel state information (CSI) and payload data among the concerned BSs. However, it requires high capacity backhaul for payload exchange between BSs, which is a cost bottleneck especially in dense small cell networks. On the other hand, the coordinated MIMO is a more cost effective technique as it only requires the exchange of real-time CSIs among the BSs to perform joint precoding. Many MIMO precoding optimization algorithms have been proposed for coordinated MIMO. For example, in [4], a WMMSE algorithm is proposed to find a stationary point of the weighted sum-rate maximization problem for multi-cell downlink systems. In [5], [6], [7], the authors proposed polite water-filling method for precoding optimization in B-MAC interference networks based on the duality principle of interference networks. Although the coordinated MIMO requires smaller backhaul capacity, the overall performance is usually much lower than that of CoMP. Recently, there have been some works conducted on multi-cell coordination with consideration of backhaul limitation. In [9], a distributed and hierarchical solution of joint beamforming and power allocation was proposed to maximize the worst-user SINR in time-division-duplex (TDD) multicell downlink systems where only limited inter-cell information exchange is permitted. In [9], random matrix theory is leveraged to design a distributed joint beamforming and power control algorithm that only requires statistical information. Such design reduces the amount of control signaling over the backhaul. An interesting question is, can we achieve the CoMP gain with reduced backhaul bandwidth consumption? We show that is possible for media streaming applications by using a novel cache-induced opportunistic CoMP scheme proposed in this paper. Specifically, we can opportunistically transform the interference network into a CoMP broadcast channel by caching a portion of the media files at the BSs. As a result, there are two transmission modes at the physical layer, namely, the CoMP mode and the coordinated MIMO mode, depending on the cache state at the BSs. If the payload data accessed by each user exists in the cache of the BSs, the BSs can engage in CoMP and therefore, enjoy a large performance gain without consuming the backhaul bandwidth. Otherwise, coordinated MIMO is employed at the BSs to serve the users. Hence, there is a cache-induced topology change in the physical layer (dynamic CoMP opportunity) of the MIMO interference network. As such, a MIMO interference network employing the cache-induced opportunistic CoMP is called a cached MIMO interference network in this paper. With high capacity caches at the BSs and a proper caching strategy, the opportunity of CoMP in the cached MIMO interference network can be very large and thus the proposed solution will have a significant gain over the coordinated MIMO scheme with even smaller backhaul consumption. Note that in the proposed solution, the reduced backhaul consumption is due to the reduced payload data transmission over the backhaul. The payload data transmission consumes much more backhaul bandwidth than the exchange of control signaling because the former needs to be done on a per-symbol basis but the latter needs to be done on a per frame basis. Hence the backhaul
saving of the proposed solution is much more significant compared to those only reduce the control signaling in the backhaul [3], [9]. Since the cost of hard disks is much lower than the cost of optical fiber backhaul, the proposed solution is very cost effective.

The performance of the proposed solution depends heavily on the dynamic caching strategy (which affects the opportunity of CoMP) and the MIMO precoding design. We study a mixed-timescale joint optimization of MIMO precoding and cache control in cached MIMO interference networks to minimize the average sum transmit power subject to fixed data rate constraints for all users. The role of cache control is to create more CoMP opportunities and is adaptive to long-term popularity of the media files (long-term control). The role of MIMO precoding optimization is to exploit the CoMP opportunities (induced by the cache) to guarantee the individual rate constraints for each user. As such, it is adaptive to the instantaneous CSI and the cache state at the BSs. There are several first order technical challenges to be addressed.

- **Limited Cache Size**: The performance gain of the proposed scheme depends heavily on the CoMP opportunity, which in turn depends on the cache size and cache strategy. The BSs usually do not have enough cache to store all the media files. As will be shown in Example 1 when brute force caching is used, even if a significant portion of the media files are cached at BSs, the CoMP opportunity can still be very small and this is highly undesirable.

- **Non-Convex Stochastic Optimization**: The mixed-timescale joint optimization of MIMO precoding and cache control is a non-convex stochastic optimization problem and the complexity of finding the optimal solution is extremely high. For example, the short-term MIMO precoding optimization in the interference networks is well known to be a difficult non-convex problem. Furthermore, the objective function for long-term cache control has no closed form expression because the short-term precoding problem has no closed form solution and the popularity of the media files is in general unknown.

- **Complex Coupling between Cache Control and Precoding Optimization**: Caching has been widely used in fixed line P2P systems [10] and content distribution networks (CDNs) [11]. In [12], a FemtoCaching scheme has also been proposed for wireless systems. However, these schemes do not consider cache-induced opportunistic CoMP among the BSs. Hence, the cache control in the above works is independent of the physical layer and is fundamentally different from our case where the cache control and physical layer are coupled together. In our case, the cache control will affect the physical layer dynamics seen by precoding optimization due to different CoMP opportunities. On the other hand, the short-term precoding strategy adopted in the physical layer will also affect the cache control due to a different cost-reward dynamic.

To address the above challenges, we first propose a novel cache data structure called *MDS-coded random cache* which can significantly improve the probability of CoMP. We then exploit the timescale separations of the optimization variables to decompose the stochastic optimization problem into a short-term precoding problem and a long-term stochastic cache control problem. We generalize the WMMSE approach in [4] to find a stationary point for the short-term precoding problem. To solve the long-term cache control problem, we first show that despite the non-convexity in the short-term precoding problem, there is a hidden convexity in the long-term stochastic cache control problem. We propose a stochastic-subgradient-like iterative solution and show that it converges to the optimal solution of this long-term stochastic optimization problem. The proposed solution has low complexity and does not require explicit knowledge of the popularity of the media files. Finally, we illustrate with simulations that the proposed solution achieves significant gain over various baselines under the consideration of overhead in the backhaul.

**Notations**: The superscript $(\cdot)^\dagger$ denotes Hermitian. The notation $1 \cdot$ denote the indication function such that $1 (E) = 1$ if the event $E$ is true and $1 (E) = 0$ otherwise. The notation $[\mathbf{A}]_{i,j}$ represents the element at the $i$-th row and $j$-th column of a matrix $\mathbf{A}$. For a square matrix $\mathbf{A}$, $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$ and $\mathbf{A} \succeq 0$ means that $\mathbf{A}$ is positive semidefinite. The notation $[a_k]_{k=1,\ldots,K}$ denote a $K \times 1$ vector whose $k$-th element is $a_k$.

## II. System Model

In this section, we introduce the architecture of the cached MIMO interference networks, the physical layer (opportunistic CoMP) and the MDS-coded random cache scheme that supports opportunistic CoMP.

### A. System Architecture of Cached MIMO Interference Networks

The architecture of the cached MIMO interference network is illustrated in Fig. 1. There are $L$ media files on the media server. The size of the $l$-th media file is $F_l$ bits and the streaming rate is denoted by $\mu_l$ (bits/s). There are $K$ users streaming media files from the media server via a radio access network (RAN) consisting of $K$ BSs and each BS is associated with one user. Each BS is equipped with $N_T \geq 2$ antennas and each user is equipped with $N_R$ antennas. The index of the media file requested by the $k$-th user is denoted by $\pi_k$. Define $\pi = \{\pi_1, \ldots, \pi_K\}$ as the user request profile (URP). We have the following assumption on URP.

**Assumption 1 (URP Assumption)**. The URP $\pi (t)$ is a slow ergodic random process (i.e., $\pi (t)$ remains constant for a large number of time slots) according to a general distribution.

The media packets (payload data) requested by the $k$-th user are delivered to the $k$-th BS from the media gateway via backhaul as illustrated in Fig. 1. Moreover, each BS is equipped with a cache of size $B_C$ bits. In this paper, the time is partitioned into time slots indexed by $t$ with duration $\tau$.

\footnote{For clarity, we consider the case where each BS is only associated with one user. However, the solution framework can be easily extended to the case with multiple users per BS.}
B. Benefits of Caching in MIMO Interference Networks

The RAN is the performance bottleneck of the system. Without caching at each BS, the RAN forms a MIMO interference network and the performance is limited by the inter-cell interference between the BSs. In this section, we propose a cache-induced opportunistic CoMP which can opportunistically use the cached media packets at the BSs to transform an interference network into a CoMP broadcast channel as illustrated in Fig. 2(b). The impact of caching at BS on the physical layer is summarized by the cache state defined as \( S \in \{ 0, 1 \} \), where \( S = 1 \) means that \( \forall k \), the current payload data requested by user \( k \) is in the cache of all the \( K \) BSs and thus it is possible for the BSs to cooperatively transmit the payload data to the users; and \( S = 0 \) means that the \( k \)-th user can only be served by the \( k \)-th BS. Hence, there are two transmission modes depending on the cache state \( S \), namely the CoMP mode \(( S = 1)\) and the coordinated MIMO mode \(( S = 0)\). Fig. 2 illustrates two examples of the data flows under different cache states \( S \) (or transmission modes). As illustrated in Fig. 2 when \( S = 1 \), the RAN can enjoy significant spatial multiplexing gain \([13]\) due to cache-induced CoMP transmission and the gain is achieved without expensive backhaul\(^2\).

We consider an OFDM based system where the wireless link between each BS and user consists of \( M \geq K \) orthogonal subcarriers. Let \( H_{m,k,n} \in \mathbb{C}^{N_t \times N_r} \) denote the channel matrix between user \( k \) and BS \( n \) on subcarrier \( m \). We have the following assumption on the global CSI \( \mathbf{H} \).

**Assumption 2 (Channel Assumption).** \( H_{m,k,n}(t) \) remains constant within a time slot but is i.i.d. w.r.t. time slot index \( t \). Specifically, \( H_{m,k,n}(t) \) has i.i.d. complex Gaussian entries of zero mean and variance \( g_{k,n} \).

The variance \( g_{k,n} \) is usually used to model the path gain between BS \( n \) and user \( k \). Note that we do not require that the channel \( H_{m,k,n}(t) \) is i.i.d. w.r.t. subcarrier index \( m \). We consider a centralized optimization scheme in which a central node computes all the control variables and then transmits them to the BSs. We assume that the central node has the knowledge of the global CSI \( \mathbf{H} \).

Then we elaborate the two transmission modes in the proposed cache-induced opportunistic CoMP.

**Coordinated MIMO Mode:** If \( S = 0 \), the \( k \)-th user can only be served by the \( k \)-th BS. We consider linear precoding and MMSE receiving for inter-cell interference cancellation. The received signal for user \( k \) on subcarrier \( m \) can be expressed as:

\[
y_{m,k} = H_{m,k,k}V_{m,k} s_{m,k} + \sum_{n \neq k} H_{m,k,n} V_{m,n} s_{m,n} + z_{m,k},
\]

where \( s_{m,k} \in \mathbb{C}^{d_{m,k}} \sim \mathcal{CN}(0, I) \) and \( d_{m,k} \) are respectively the data vector and the number of data streams for user \( k \) on subcarrier \( m \); \( V_{m,k} \in \mathbb{C}^{N_t \times d_{m,k}} \) is the precoding matrix for user \( k \) on subcarrier \( m \); and \( z_{m,k} \in \mathbb{C}^{N_r} \sim \mathcal{CN}(0, I) \) is the AWGN noise vector. The MMSE receiver for user \( k \) on subcarrier \( m \) is given by

\[
U_{m,k} = \left( \Omega_{m,k} + H_{m,k,k} V_{m,k} V_{m,k}^\dagger H_{m,k,k}^\dagger \right)^{-1} H_{m,k,k} V_{m,k},
\]

where \( \Omega_{m,k} = I + \sum_{n \neq k} H_{m,k,n} V_{m,n} V_{m,n}^\dagger H_{m,k,n}^\dagger \) is the interference-plus-noise covariance matrix of user \( k \) on subcarrier \( m \). Then for given CSI \( \mathbf{H} \), cache state \( S = 0 \) and precoding matrices \( \mathbf{V} = \{ V_{m,k} : \forall m, k \} \), the data rate (bps) of user \( k \) is

\[
R_k(\mathbf{H}, \mathbf{V}) = \frac{B_W}{M \ln 2} \sum_{m=1}^{M} r_{m,k},
\]

where \( B_W \) is the bandwidth of the system; and

\[
r_{m,k} = \log \left( I + H_{m,k,k} V_{m,k} V_{m,k}^\dagger H_{m,k,k}^\dagger \Omega_{m,k}^{-1} \right).
\]

The corresponding sum transmit power is

\[
P(\mathbf{V}) = \sum_{k=1}^{K} \sum_{m=1}^{M} \text{Tr} \left( V_{m,k} V_{m,k}^\dagger \right).
\]

**CoMP Mode:** If \( S = 1 \), the \( K \) users are served using CoMP between the BSs. Similarly, we consider linear precoding and
MMSE receiving for interference cancellation. The received signal for user \( k \) on subcarrier \( m \) can be expressed as:

\[
\hat{y}_{m,k} = \hat{H}_{m,k} \hat{V}_{m,k} s_{m,k} + \sum_{n \neq k} \hat{H}_{m,n} \hat{V}_{m,n} s_{m,n} + z_{m,k},
\]

where \( \hat{H}_{m,k} = [\hat{H}_{m,k,1}, ..., \hat{H}_{m,k,K}] \in \mathbb{C}^{N_{R} \times KN_{R}} \) is the composite channel matrix between all the BSs and user \( k \); \( s_{m,k} \in \mathbb{C}^{d_{m,k}} \sim \mathcal{CN}(0, \mathbf{I}) \) and \( d_{m,k} \) are respectively the data vector and the number of data streams for user \( k \) on subcarrier \( m \); and \( \hat{V}_{m,k} \in \mathbb{C}^{KN_{R} \times d_{m,k}} \) is the composite precoding matrix for user \( k \) on subcarrier \( m \). In CoMP mode, the MMSE receiver for user \( k \) on subcarrier \( m \) is given by

\[
\hat{U}_{m,k} = \left( \Omega_{m,k} + \hat{H}_{m,k} \hat{V}_{m,k} \hat{V}_{m,k}^{\dagger} \hat{H}_{m,k}^{\dagger} \right)^{-1} \hat{H}_{m,k} \hat{V}_{m,k},
\]

where \( \Omega_{m,k} = \mathbf{I} + \sum_{n \neq k} \hat{H}_{m,k} \hat{V}_{m,n} \hat{V}_{m,n}^{\dagger} \hat{H}_{m,k}^{\dagger} \) is the interference-plus-noise covariance matrix of user \( k \) on subcarrier \( m \). Then for given CSI \( \hat{H} \), cache state \( S = 1 \) and precoding matrices \( \hat{V} = \{ \hat{V}_{m,k} : \forall m, k \} \), the data rate (bps) of user \( k \) is

\[
R_{k}(\hat{H}, \hat{V}) = \frac{B_{W}}{M \ln 2} \sum_{m=1}^{M} \overline{r}_{m,k},
\]

where

\[
\overline{r}_{m,k} = \log \left| \mathbf{I} + \hat{H}_{m,k} \hat{V}_{m,k} \hat{V}_{m,k}^{\dagger} \hat{H}_{m,k}^{\dagger} \Omega_{m,k}^{-1} \right|.
\]

The corresponding sum transmit power is

\[
\overline{P}(\hat{V}) = \frac{K}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} \text{Tr} \left( \hat{V}_{m,k} \hat{V}_{m,k}^{\dagger} \right).
\]

The choice of the number of data streams \( \{d_{m,k}\} \) and \( \{\hat{d}_{m,k}\} \) is an important problem. In [4], the number of data streams is treated as a system parameter and there is no discussion about how to choose this parameter. In this paper, we show that it will not lose “optimality” to choose the number of data streams for the coordinated MIMO mode to be

\[
d_{m,k} = d \triangleq \min (N_{T}, N_{R}), \forall m, k,
\]

in the CoMP mode to be

\[
\hat{d}_{m,k} = \hat{d} \triangleq \min (KN_{T}, N_{R}), \forall m, k.
\]

The overall performance gain of the cache-induced opportunistic CoMP depends heavily on the probability of \( S = 1 \). In this section, we propose a novel MDS-coded random cache scheme which makes the best use of the BS cache to increase \( \Pr[S = 1] \). We first use an example to show that, with a naive cache scheme, the CoMP opportunity \( \Pr[S = 1] \) can be very small even if a significant portion of the media files are stored at the BS cache.

Example 1 (Brute-force cache data structure). Suppose that there are \( K = 4 \) BS-user pairs and \( L = 4 \) media files with equal size of \( F \) bits. The \( k \)-th media file is requested by the \( k \)-th user. Each BS randomly stores half of the media packets (i.e., \( 0.5F \) bits) for each media file. Then the probability that the packets requested by a single user are in the cache of all the BSs is \( 0.5^{K} = 0.0625 \). However, the probability of \( S = 1 \) is only \( 0.0625^{K} < 0.00002 \).

Hence, a more intelligent cache scheme is needed. In the following, we propose a novel MDS-coded random cache data structure which can significantly improve the probability of CoMP.

**Cache Data Structure:** Each media file is divided into segments. Specifically, each segment of the \( l \)-th media file contains \( T_{S}M_{l}\tau \) bits, where \( T_{S} \gg 1 \) is some integer, and it is encoded into \( T_{S}M_{l}\tau \) parity bits using an ideal MDS rateless code as illustrated in Fig. 3(a). An MDS rateless code generates an arbitrarily long sequence of parity bits from...
Figure 4: An example on how to generate the cache state \( S \). At each time slot, the cache state \( S \) is first generated by a random cache state generator at the central node (as illustrated in Fig. 4) and then it is broadcast to the BSs. In this example, we assume that \( T_S = 8 \) and \( q_{\text{min}} = 3/8 \). The index set \( T_S = \{1, 3, 6\} \). The red box contains a time interval of \( T_S = 8 \) consecutive time slots. It is easy to see that for any time interval of \( T_S = 8 \) consecutive time slots, there are \( q_{\text{min}} T_S \) time slots with \( S = 1 \).

Figure 5: An illustration of playback buffer, reassembling buffer and media decoding at the user. Once the total number of parity bits for current segment at the reassembling buffer is equal to \( T_{S\mu_\tau} \), the whole segment is decoded at the media decoder and the reassembling buffer is cleared so that the media decoder can read the next segment from the playback buffer.

an information packet of \( L_S \) bits (\( L_S \) can be any positive integer), such that if the decoder obtains any \( L_S \) parity bits, it can recover the original \( L_S \) information bits. In practice, the MDS rateless code can be implemented using Raptor codes \([4]\) at the cost of a small redundancy overhead. The cache at each BS stores \( q T_{S\mu_\tau} \) parity bits for every segment of the \( l \)-th media file as illustrated in Fig. 3(b), where \( q_l \in \{0, \frac{1}{T_S}, ..., \frac{T_S-1}{T_S}, 1\} \) is called the cache control variable. Such a cache data structure is more flexible than the brute-force cache data structure in Example 1 in the sense that we can control when to use the cached data.

**Random Cache Usage:** The cache usage is determined by the cache state \( S \) as illustrated in Fig. 3(c,d). Using the above MDS-coded cache data structure, we can actively control the cache state \( S \) for each time slot. The key to increasing the probability of CoMP in the system is to align the transmissions of the cached data for different users as much as possible. Specifically, for given cache control vector \( q = [q_1, ..., q_T]^T \) and URP \( \pi \), let \( q_{\text{min}} = \min_{1 \leq k \leq K} \{q_{\pi_k}\} \). Then conditioned on a given \( q \) and \( \pi \), the cache state \( S \) is generated by a random cache state generator at the central node (as illustrated in Fig. 4) using the following method. First, time is divided into frames where each frame contains \( T_S \) time slots. Then the central node randomly generates an index set \( T_S \subseteq \{1, 2, ..., T_S\} \) such that \( |T_S| = q_{\text{min}} T_S \). Note that for given \( q \) and \( \pi \), \( T_S \) is only generated for once and it remains constant until \( q \) and \( \pi \) changes to a new value. Suppose that the current time slot \( t \) is the \( i \)-th time slot in the current frame. Then if \( i \in T_S \), we let \( S(t) = 1 \); and otherwise, we let \( S(t) = 0 \). Finally, the central node broadcasts the cache state \( S(t) \) to the BSs. Fig. 4 gives an example of how to generate the cache state \( S \) for each time slot. It can be seen that for any time interval of \( T_S \) consecutive time slots, there are \( q_{\text{min}} T_S \) time slots with \( S = 1 \) and \( (1 - q_{\text{min}}) T_S \) time slots with \( S = 0 \), as illustrated in Fig. 4.

The BSs decide when to do MIMO cooperation according to the generated cache state \( S \). If \( S = 1 \), the BSs employ CoMP to jointly transmit the cached parity bits to the \( K \) users without consuming the backhaul bandwidth as illustrated in Fig. 3(d). Otherwise, BS \( k \) obtains the parity bits requested by user \( k \) from the backhaul and transmits them to user \( k \) using the coordinated MIMO transmission mode as illustrated in Fig. 3.

**Media Decoding at Each User:** At user \( k \), the process of receiving and decoding a segment of the \( l \)-th media file, where \( l = \pi_k \), is summarized as follows. There is a playback buffer and a media decoder at the user terminal. The media decoder has a reassembling buffer as illustrated in Fig. 5. At each time slot, if \( S = 1 \), user \( k \) receives \( \mu_\tau \) cached parity bits of current segment from all BSs using CoMP and stores them in the playback buffer. If \( S = 0 \), user \( k \) receives \( \mu_\tau \) uncached parity bits of current segment from BS \( k \) using coordinated MIMO and stores them in the playback buffer. User \( k \) keeps receiving parity bits from the RAN until the total number of received parity bits for current segment is equal to \( T_{S\mu_\tau} \). Then, user \( k \) starts to receive the next segment from the RAN. On the other hand, the media decoder keeps reading parity bits from the playback buffer at a constant playback rate (which is equal to \( \mu_\tau \)) and storing them in the reassembling buffer until the total number of parity bits for current segment at the reassembling buffer is equal to \( T_{S\mu_\tau} \). Then the whole segment is decoded and the reassembling buffer is cleared so that the media decoder can read the next segment from the playback buffer. Clearly, it takes \( T_S \) time slots for user \( k \) to receive all the \( T_{S\mu_\tau} \) parity bits and the number of time slots with \( S = 1 \) is \( q_{\text{min}} T_S \). Hence, user \( k \) receives a total number of \( q_{\text{min}} T_{S\mu_\tau} \) cached parity bits for each segment, which is feasible (i.e., there is no BS cache underflow) since each BS stores \( q T_{S\mu_\tau} \) parity bits for every segment of the \( l \)-th media file.

The following example illustrates the advantage of MDS-coded random cache scheme.

**Example 2 (Advantage of MDS-coded random cache).** Consider the setup in Example 1. We have \( q_{\pi_k} = 0.5, \forall \pi_k \) and thus the probability of \( S = 1 \) is 0.5, which is much larger than that of the brute-force caching scheme in Example 1 (\( < 0.00002 \)). Compared with the brute-force caching scheme in Example 1, the probability of CoMP transmission (\( S = 1 \)) under the proposed MDS-coded random cache is \( \min_{k} \{q_{\pi_k}\} \) versus \( 2 \).

\(^3\)In general, the arrival packets from the RAN is burst and the playback buffer is used to maintain a constant playback rate at the media decoder.

\(^4\)For the brute-force caching scheme, the probability that the packets requested by user \( k \) are in the cache of all the BSs is \( q_k^K \). For \( k \neq 1 \), whether the packets requested by user \( k \) is in the BS caches is independent of whether the packets requested by user \( l \) is in the BS caches. Hence, the probability of \( S = 1 \) is \( \prod_{k=1}^{K} q_k^K \).
$\Pi_{k=1}^{K} q_{\pi k}$. This represents a first order improvement in the opportunity of CoMP gain. Yet, there is a fundamental tradeoff between the performance gain and the BS cache size. Intuitively, the more popular the media file is, the larger portion of its parity bits should be stored in the BS cache to increase the CoMP probability. Hence, the value of $q$ must be carefully controlled to achieve the best tradeoff among performance and the BS cache size. As such, the cache control variable is parameterized by the vector $q = [q_1, ..., q_L]^T$.

III. Mixed Timescale Precoding and Cache Control

In this section, we formulate a mixed-timescale optimization problem for media streaming under cache-induced opportunistic CoMP. The control variables are partitioned into long-term and short-term control variables. The long-term control variables (cache control variables $q$) are adaptive to the distribution of the URP $\pi$ to induce CoMP opportunity. The short-term control variables (precoding matrices $V, V$) are adaptive to the instantaneous cache state $S$ and CSI $H$ to exploit the opportunistic CoMP gain and to guarantee the QoS requirements of the users for given $q$ and $\pi$.

A. Problem Formulation

For convenience, let $V(\pi, H) = \{V_{m,k}(\pi, H) : \forall m, k\}$ denote all precoding matrices under URP $\pi$, CSI $H$, and cache state $S = 0$; and let $\tilde{V}(\pi, H) = \{\tilde{V}_{m,k}(\pi, H) : \forall m, k\}$ denote all precoding matrices under URP $\pi$, CSI $H$, and cache state $S = 1$. Define $V = \{V(\pi, H), \tilde{V}(\pi, H) : \forall \pi, H\}$ as the collection of precoding matrices for all possible URP,CSI, and cache state combinations $\{\pi, H, S\}$. Then for given set of control variables $(q, V)$ and URP $\pi$, the average sum transmit power is given by

$$\bar{P}_{\pi}(q, V) = (1 - \min_{k} q_{\pi k}) E[|P(V(\pi, H))|] + \min_{k} q_{\pi k} E\left[|\tilde{P}(\tilde{V}(\pi, H))|\right].$$  \hspace{1cm} (6)

For convenience, define the feasible sets for cache control $q$, coordinated precoding $V(\pi, H)$ and CoMP precoding $V(\pi, H)$ respectively as

$$D_q = \left\{ q : q_l \in [0,1], \forall l, \text{ and } \sum_{l=1}^{L} F_l q_l \leq B_C \right\},$$

$$D_v(\pi, H) = \left\{ V : R_k(H, V) \geq \mu_{\pi_k}, \forall k \right\},$$

$$D_{\tilde{V}}(\pi, H) = \left\{ \tilde{V} : \tilde{R}_k(H, \tilde{V}) \geq \mu_{\pi_k}, \forall k \right\}. \hspace{1cm} (7)$$

In $D_q$, $\sum_{l=1}^{L} F_l q_l \leq B_C$ is the BS cache size constraint used to avoid BS cache overflow. Note that $q_l$ is relaxed to be a real number in $[0,1]$. This relaxation has little effect on the performance when $T_S \gg 1$. In $D_v(\pi, H)$ ($D_{\tilde{V}}(\pi, H)$), $R_k(H, V)\geq \mu_{\pi_k}$ ($\tilde{R}_k(H, \tilde{V})\geq \mu_{\pi_k}$) is the instantaneous rate constraint for user $k$ under URP $\pi$ and cache state $S = 0$ ($S = 1$). In media streaming applications, the playback process at user $k$ can be modeled by a playback queue with random arrival (from the BS via the RAN) and deterministic departure as illustrated in Fig. 5. The media streaming QoS can be represented by playback interruption probability, which is the same as the probability of playback buffer being empty as indicated in Fig. 6. Since the departure process is deterministic with constant rate $\mu_{\pi_k}$, the instantaneous rate constraint $R_k (H, V) \geq \mu_{\pi_k}$ (when $S = 0$) and $\tilde{R}_k (H, \tilde{V}) \geq \mu_{\pi_k}$ (when $S = 1$) in (7) essentially guarantee that the playback process is free from interruption.

Then the joint cache and power control problem is formulated as:

$$P : \min_{q \in D_q, V} \mathbb{E}\left[\bar{P}_{\pi}(q, V)\right]$$

s.t. $V(\pi, H) \in D_v(\pi, H), \tilde{V}(\pi, H) \in D_{\tilde{V}}(\pi, H), \text{ w.p.1}$, \hspace{1cm} (8)

where the expectation is taken w.r.t. the distribution of $\pi$. Note that in constraint (8), the instantaneous rate requirement is satisfied with probability one because it is impossible to guarantee a fixed data rate for all realizations of $H$.

In general, problem $P$ may not even be feasible (i.e. $\mathbb{E}[\bar{P}_{\pi}(q, V)]$ is not bounded when constraint (8) is satisfied). However, the following proposition ensures that $P$ is feasible.

Proposition 2 (Feasibility of $P$). There exists $V = \{V(\pi, H), \tilde{V}(\pi, H) : \forall \pi, H\}$ such that $V(\pi, H) \in D_v(\pi, H), \tilde{V}(\pi, H) \in D_{\tilde{V}}(\pi, H)$ with probability one and $\mathbb{E}[\bar{P}_{\pi}(q, V)]$ is bounded.

Please refer to Appendix B for the proof.

B. Problem Decomposition

Problem $P$ is a non-convex stochastic optimization problem. We first decompose it into simpler subproblems. For convenience, define

$$H_F \triangleq \left\{ H : \text{Tr} \left( H_{m,k,k} H_{m,k,k}^\dagger \right) > 0, \forall m, k \right\}. \hspace{1cm} (9)$$

According to the analysis in Appendix B for any $\pi$ and $H \in H_F$, there exists $V, \tilde{V}$ such that $V \in D_v(\pi, H), \tilde{V} \in D_{\tilde{V}}(\pi, H)$ and $P(V), \tilde{P}(\tilde{V})$ are bounded. Moreover, we have $\Pr[H \in H_F] = 1$. Then by exploiting the timescale separations of the optimization variables, problem $P$ can be decomposed into the following families of subproblems. Subproblem 1 (Short-term Coordinated MIMO Precoding for given $\pi$, $H \in H_F$ and $S = 0$):

$$P_S(\pi, H) : \min_{V} P(V), \text{ s.t. } V \in D_v(\pi, H).$$
Subproblem 2 (Short-term CoMP Preceding) for given $\pi$, $H \in \mathcal{H}_F$ and $S = 1$:

$$\tilde{P}_S (\pi, H) : \min_{\tilde{V}} \tilde{P} \left( \tilde{V} \right), \text{s.t. } \tilde{V} \in D_\pi (\pi, H).$$

Subproblem 3 (Long-term Cache Control) for given $\mathcal{V}$:

$$P_L (\mathcal{V}) : \min_{\mathcal{Q} \in D_\pi} \psi (\mathcal{Q}, \mathcal{V}) = E \left[ P_\pi (\mathcal{Q}, \mathcal{V}) \right]. \quad (10)$$

The relationship between problem $P$ and the subproblems is summarized as follows. For given $\pi, H \in \mathcal{H}_F$, let $V^* (\pi, H)$ be the optimal solution of $P^* (\pi, H)$ and $V^* (\pi, H)$ be the optimal solution of $P_S (\pi, H)$. For given $\pi, H \notin \mathcal{H}_F$, let $V^* (\pi, H) = \{ V^*_{m,k} (\pi, H) = 0 : \forall m,k \}$ and $V^* (\pi, H) = \{ V^*_{m,k} (\pi, H) = 0 : \forall m,k \}$. Let $q^*$ be the optimal solution of $P_L (\mathcal{V})$, where $V^* = \{ V^* (\pi, H), \tilde{V}^* (\pi, H) : \forall \pi, H \}$. Then $(q^*, V^*)$ is the optimal solution of $P$.

The above three subproblems are still non-trivial. Although the gradient projection (GP) method is usually used to find a stationary point for a constrained non-convex problem, it cannot be applied to solve $P_S (\pi, H)$ because $P_\pi (\pi, H)$ is a non-convex set and calculating the projection of $V$ on the feasible set $D_\pi (\pi, H)$ is also a non-convex problem. Similar observations can also be made for $P_S (\pi, H)$. In Section IV, we generalize the WMSE approach in [4] to obtain a polynomial complexity algorithm which converges to a stationary point of the short-term precoding problem $P_S (\pi, H)$ or $P_S (\pi, H)$. In Section IV, we exploit the hidden-convexity and propose a robust stochastic subgradient algorithm to solve $P_L (\mathcal{V})$.

IV. Short-term Preceding Solutions for $P_S (\pi, H)$

and $P_S (\pi, H)$

Problem $P_S (\pi, H)$ is a sum power minimization problem under individual rate constraints in parallel interference networks. In [4], a WMSE algorithm was proposed to find a stationary point for the weighted sum-rate maximization problem (WSRMP) in MIMO interfering broadcast channels under per-Bs power constraints. In the following, the WMSE algorithm is generalized to solve $P_S (\pi, H)$ and $P_S (\pi, H)$.

A. An Equivalent Problem under Weighted MSE Constraint

Consider a sum power minimization problem under individual weighted MSE constraints:

$$\min_{\mathcal{W}, \mathcal{U}, \mathcal{V}} P \left( \mathcal{V} \right) = \sum_{k=1}^{K} \sum_{m=1}^{M} \text{Tr} \left( V_{m,k} V_{m,k}^\dagger \right) \quad (11)$$

$$\text{s.t. } \sum_{m=1}^{M} \left( \text{Tr} \left( W_{m,k} E_{m,k} \right) - \log |W_{m,k}| \right) \leq d - \bar{p}_{\pi_k}, \forall k,$$

where $\mathcal{W} = \{ W_{m,k} \geq 0 : \forall m,k \}$ is a set of weight matrices; $\mathcal{U} = \{ U_{m,k} : \forall m,k \}$ is the set of all receiving matrices; $\bar{p}_{\pi_k} = \frac{\mu_{\pi_k}}{2} \ln 2$ and

$$E_{m,k} = \left( I - U_{m,k}^\dagger H_{m,k} V_{m,k} \right) \left( I - U_{m,k}^\dagger H_{m,k} V_{m,k} \right)^\dagger + U_{m,k}^\dagger \Omega_{m,k} U_{m,k}, \quad (12)$$

is the MSE matrix of user $k$ on subcarrier $m$. The following theorem establishes the equivalence between $P_S (\pi, H)$ and Problem (11).

Theorem 1 (Equivalence between $P_S (\pi, H)$ and (11)). For given $\pi, H \in \mathcal{H}_F$, let $(W^*, U^*, V^*)$ denote the optimal solution of Problem (11). Then $V^*$ is also the optimal solution of $P_S (\pi, H)$.

Please refer to Appendix C for the proof.

Hence we only need to solve Problem (11), which is convex in each of the optimization variables $W, U, V$. We can use the block coordinate descent method to solve (11). First, for fixed $V, U$, the optimal $W$ is given by $W_{m,k} = E_{m,k}, \forall m, k$. Second, for fixed $V, W$, the optimal $U$ is given by the MMSE receiver in (1). Finally, for fixed $U, W$, Problem (11) is a convex quadratic optimization problem which can be solved using the Lagrange dual method as will be elaborated in the next subsection.

B. Lagrange dual method for solving Problem (11) with fixed $U, W$

The Lagrange function of Problem (11) with fixed $U, W$ is given by

$$L (\lambda, \mathcal{V}, U, W) = \sum_{k=1}^{K} \sum_{m=1}^{M} \text{Tr} \left( V_{m,k} V_{m,k}^\dagger \right) + \sum_{k=1}^{K} \lambda_k \left( \sum_{m=1}^{M} \left( \text{Tr} \left( W_{m,k} E_{m,k} \right) - \log |W_{m,k}| \right) - d - \bar{p}_{\pi_k} \right),$$

where $\lambda = \{ \lambda_k \}_{k=1}^{K} \in \mathbb{R}^K_+$ is the Lagrange multiplier vector. The dual function of Problem (11) with fixed $U, W$ is

$$J (\lambda) = \min_{\mathcal{V}} L (\lambda, \mathcal{V}, U, W). \quad (13)$$

The minimization problem in (13) can be decomposed into $M$ independent problems as

$$\min_{\mathcal{V}_{m,k}} \sum_{k=1}^{K} \text{Tr} \left( V_{m,k} V_{m,k}^\dagger \right) + \sum_{k=1}^{K} \lambda_k \frac{\text{Tr} \left( V_{m,k} U_{m,k} U_{m,k}^\dagger H_{m,k} \right)}{M}, \forall m. \quad (14)$$

For fixed $\lambda$, Problem (14) has a closed-form solution given by

$$V^*_{m,k} (\lambda) = \left( \sum_{n=1}^{K} \lambda_n \frac{H_{m,n}^\dagger U_{m,n} W_{m,n} U_{m,n}^\dagger H_{m,n,k}}{M} + I \right)^{-1} \times \frac{\lambda_k}{M} H_{m,k}^\dagger U_{m,k} W_{m,k}, \forall k. \quad (15)$$

Since Problem (11) with fixed $U, W$ is a convex quadratic optimization problem, the optimal solution is given by $V^* (\lambda^*) = \{ V^*_{m,k} (\lambda^*) : \forall m, k \}$, where $\lambda^*$ is the optimal solution of the dual problem

$$\max_{\lambda} J (\lambda), \text{s.t. } \lambda \geq 0. \quad (16)$$
The dual function $J(\lambda)$ is concave and it can be verified that
\[
\left[\sum_{m=1}^{M} (\text{Tr}(W_{m,k}E_{m,k}(\lambda)) - d + \pi_{k}) \right]_{k=1,...,K}
\]
is a subgradient of $J(\lambda)$, where $E_{m,k}(\lambda)$ is obtained from \[12\] with $V_{m,k} = V_{m,k}^*$. Hence, the standard subgradient based methods such as the subgradient algorithm in \[16\] or the Ellipsoid method in \[17\] can be used to solve the optimal solution $\lambda^*$ of the dual problem in \[16\].

C. Overall Algorithm for Solving $P_S(\pi,H)$ and $\tilde{P}_S(\pi,H)$

The overall algorithm (named Algorithm SP) for solving $P_S(\pi,H)$ is summarized in Table I. Note that using the Matrix Inversion Lemma, it can be shown that $E_{m,k} = \left(I - U_{m,k}^\dagger H_{m,k,k} U_{m,k} \right)^{-1}$ if $U_{m,k}$ is the MMSE receiver given in the step 1 of Algorithm SP. Hence in step 2, we let $W_{m,k} = \left(I - U_{m,k}^\dagger H_{m,k,k} V_{m,k} \right)^{-1}$. The following theorem shows that Algorithm SP converges to a stationary point of $P_S(\pi,H)$.

**Theorem 2 (Convergence of Alg. SP).** For given $\pi,H \in \mathcal{H}_F$, any limit point $(\tilde{V}^*, \tilde{U}^*, \tilde{W}^*)$ of the iterates generated by Algorithm SP is a stationary point of Problem \[17\], and the corresponding $\tilde{V}^*$ is a stationary point of $P_S(\pi,H)$.

Please refer to Appendix D for the proof.

Problem $\tilde{P}_S(\pi,H)$ is a sum power minimization problem under individual rate constraints in parallel broadcast channel, which can be viewed as a parallel interference network with the cross link channel equal to the direct link channel. Hence, Algorithm SP can also be used to find a stationary point $\tilde{V}^*(\pi,H)$ of $\tilde{P}_S(\pi,H)$ by replacing $H_{m,k,n}, n, V_{m,k}, U_{m,k}, \Omega_{m,k}$ respectively with $H_{m,k}, V_{m,k}, U_{m,k}, \Omega_{m,k}$ and using the following initial point:

\[
[\tilde{V}_{m,k}]_{i,j} = \begin{cases} 
[V_{m,k}(\pi,H)]_{i-(k-1)N_T,j}, & i \in I_k, j \in [1,d] \\
0, & \text{otherwise}
\end{cases}
\]

for all $m,k$, where $I_k = [(k-1)N_T + 1, kN_T]$; and $\tilde{V}^*(\pi,H) = \{V_{m,k}(\pi,H), \forall m, k\}$ is the stationary point of $\tilde{P}_S(\pi,H)$ found by Algorithm SP. The initial point in \[18\] is chosen to ensure that $P(\tilde{V}^*(\pi,H)) \leq P(\tilde{V}^*(\pi,H))$, which is the key to prove the convexity of the long-term cache control problem $P_L(\tilde{V}^*)$ in Lemma I.

V. LONG TERM CACHE CONTROL FOR $P_L(\tilde{V}^*)$

The following lemma shows that $P_L(\tilde{V}^*)$ is a convex stochastic optimization problem as long as $\tilde{V}^*$ is a stationary point of $\tilde{P}_S(\pi,H)$ and $\tilde{P}_S(\pi,H)$ found by Algorithm SP.

**Lemma 1 (Hidden Convexity of $P_L(\tilde{V}^*)$).** For any $\pi$ and $H \in \mathcal{H}_F$, let $\tilde{V}^*(\pi,H)$ be the stationary point of $\tilde{P}_S(\pi,H)$ found by Algorithm SP and let $\tilde{V}^*(\pi,H)$ be the stationary point of $P_S(\pi,H)$ found by Algorithm SP with the initial point given in \[18\]. For given $\pi,H \not\in \mathcal{H}_F$, let $\tilde{V}^*(\pi,H) = \{V_{m,k}(\pi,H) = 0 : \forall m,k\}$ and $\tilde{P}_S(\pi,H)$ found by Algorithm SP.

| Step 1: | Let $U_{m,k} = \left(\Omega_{m,k} + H_{m,k,k} V_{m,k} H_{m,k,k}^\dagger \right)^{-1} H_{m,k,k} V_{m,k}, \forall m,k$. |
|--------|--------------------------------------------------------------|
| Step 2: | Let $W_{m,k} = \left(I - U_{m,k}^\dagger H_{m,k,k} V_{m,k} \right)^{-1}, \forall m,k$. |
| Step 3: | Let $V_{m,k} = \tilde{V}_{m,k}(\pi,H), \forall m,k$, where $\lambda^*$ is the optimal solution of \[16\] which can be solved using, e.g., the subgradient algorithm in \[16\] or the Ellipsoid method in \[17\] with the subgradient of $J(\lambda)$ given in \[15\] and $V_{m,k}(\pi,H)$ is given in \[15\]. |

Please refer to Appendix E for the proof.

Hence, we propose a stochastic subgradient algorithm which is able to converge to the optimal solution of $P_L(\tilde{V}^*)$ without knowing the distribution of $\pi$.

For notation convenience, let $T_i$ denote the time slot when $\pi$ changes for the $i$-th time, i.e., $\pi$ remains constant for each time interval $[T_i, T_{i+1} - 1]$ and changes at the boundary of each time interval. Then the following lemma gives a noisy unbiased subgradient of the objective function $\psi(q, \tilde{V}^*)$.

**Lemma 2 (Noisy unbiased subgradient of $\psi(q, \tilde{V}^*)$).** Let $\pi^{(i)}$ denote the URP for the $i$-th time interval $[T_i, T_{i+1} - 1]$. At time slot $T_{i+1} - 1$, a noisy unbiased subgradient of $\psi(q, \tilde{V}^*)$ at $q \in D_q$, denoted by $\tilde{\psi}(\psi(i)) = \left[\frac{\partial \psi(i)}{\partial q_1}, ..., \frac{\partial \psi(i)}{\partial q_L}\right]^T$, is given by

\[
\frac{\partial \psi(i)}{\partial q_l} = 1 \left(l = \pi^{(i)}_k \right) \left( \sum_{t \in T(i)} \tilde{P}(\tilde{V}^*(\pi^{(i)}, H(t))) \tilde{\tau}(i) \tilde{\tau}(i) \right) - \sum_{t \in T(i)} P(\tilde{V}^*(\pi^{(i)}, H(t))) \tilde{\tau}(i), \forall l
\]

for all $l$, where $k$ is any index satisfying $q_{\pi_k} = \min_{1 \leq k \leq K} \{q_{\pi_k}\}$; $T(i), \tilde{T}(i)$ are any two non-empty sets of time slot index in $[T_i, T_{i+1} - 1]$.

Please refer to Appendix F for the proof.

In practical implementation, we can choose $T(i) = \{t \in [T_i, T_{i+1} - 1] : S(t) = 0\}$ and $\tilde{T}(i) = \{t \in [T_i, T_{i+1} - 1] : S(t) = 1\}$ to avoid the extra computation for the subgradient calculation, providing that the two sets are not empty.

After obtaining the noisy unbiased subgradient using \[19\], $q^{(i)}$ is updated using the following subgradient projection method

\[
q^{(i+1)} = \arg\min_{q \in D_q} \|q^{(i)} - \sigma^{(i)} \psi(i)(q^{(i)}) - q\|^2,
\]

where $\sigma^{(i)} > 0$ is the step size for the $i$-th update. The projection method in \[20\] is a convex quadratic optimization.
Table II: Algorithm LC (for Solving Problem $P_L (V^*)$)

| Initialization: Let $q^{(0)} = 0$ and $i = 0$. |
|--------------------------------------------------|
| Step 1: At time slot $t = T_{i+1} - 1$, calculate a noisy unbiased subgradient of $\psi(q^{(i)}, V^{*})$ using (19). |
| Step 2: Choose proper step size $\sigma^{(i)} > 0$ and obtain $q^{(i+1)}$ using (20). |
| Step 3: Let $i = i + 1$ and return to Step 1. |

B. Complexity in Computation and Signaling

Computation Complexity: The complexity of the precoding optimization Algorithm SP is similar to the WMMSE algorithm in [4] and is polynomial w.r.t. the number of users and antennas at each node. Please refer to [4] for details. The long term cache control only needs to do a simple subgradient projection update in (24) for each realization of $\pi$ and thus the complexity is extremely low.

Control Signaling Overhead: The short term control signaling overhead is similar to the conventional coordinated MIMO schemes and can be supported by the modern wireless systems such as LTE [19]. At each time slot, user $k$ feedbacks its direct and cross link CSI $H_{m,k,n}, \forall m,n$ to the central node. Then the central node broadcasts the cache state $S$ and precoding matrices $V$ or $\tilde{V}$ to the BSs. The long term cache control signaling between the BSs and central node is very small since $q$ is only sent to the BS once for each realization of $\pi$.

Average Backhaul Consumption: For simplicity, we assume $\mu_t = \mu_0, \forall l$. Suppose Algorithm LC converges to $q$. We first analyze the backhaul consumption due to online media streaming. For each segment of the media file requested by user $k$, there are only $(1 - \min_k \{q_{\pi_k}\}) \mu_0 T_S$ parity bits from the backhaul when the URP is $\pi$. Hence, the total average backhaul consumption due to online media streaming for $K$ users is $E[K (1 - \min_k \{q_{\pi_k}\}) \mu_0]$. For the BS cache update process, each BS needs to obtain a total number of $\sum_{l=1}^{L} q_l F_l$ parity bits from the backhaul. However, these cache updates can be done offline with much smaller average backhaul consumption compared with the backhaul consumption due to online streaming because the popularity of media files changes very slowly (e.g. new movies are usually posted on a weekly or monthly timescale). Let $T_C$ denote the interval of the cache update process, then the average backhaul consumption due to the offline BS cache update is $K \sum_{l=1}^{L} q_l F_l / T_C$. Finally, the overall average backhaul consumption (bps) is given by

$$R_B = E[K (1 - \min_k \{q_{\pi_k}\}) \mu_0] + K \sum_{l=1}^{L} q_l F_l / T_C. \quad (21)$$

Table III compares the average backhaul consumption of different schemes. Impressively, the average backhaul consumption is much smaller than the coordinated MIMO for only moderately large BS cache size $B_C$.

VII. Simulation Results

Consider a media streaming system with $L = 6$ media files and $K = 7$ BS-user pairs. The BSs are arranged as in Fig.
Table III: Backhaul consumption comparison of different schemes. The corresponding average transmit power required for QoS guarantee is also illustrated. We assume that the popularity of the media files (i.e., the distribution of $\pi$) changes every week, i.e., $T_C = 1$ week. The system setup is given in Section VII and the results are evaluated under edge user placement with a streaming rate of $\mu_l - \mu_0 = 2\text{Mbits/s}, \forall l$. The backhaul consumption under normal user placement is similar and is omitted for conciseness.

![Figure 8: Topology of a multi-cell downlink with 7 cells.](image)

![Figure 9: Objective value of $P_S(\pi, H)$ versus the number of iterations of Algorithm SP.](image)

![Figure 10: Objective value of $P_L(\mathcal{V}^*)$ versus the number of iterations of Algorithm LC.](image)

A. Convergence of Algorithm SP and LC

Consider normal user placement and set the streaming rate as $\mu_l - \mu_0 = 2\text{Mbits/s}, \forall l$. In Fig. 9 we plot the objective value $P(\mathcal{V})$ of $P_S(\pi, H)$ versus the number of iterations for a single realization of $\pi, H$. In Fig. 10 we plot the objective value $\psi(q, \mathcal{V}^*)$ of $P_L(\mathcal{V}^*)$ versus the number of realizations of $\pi$ (i.e., the number of iterations of Algorithm LC) for different BS cache size $B_C$. It can be seen that both Algorithm SP and LC quickly converge.

B. Advantage of the Proposed Solution w.r.t. Baselines

The following baselines are considered.

**Baseline 1 (Coordinated MIMO):** The BS has no cache. The physical layer reduces to the coordinated MIMO precoding scheme in Section II-B. At each time slot, the precoding matrices are obtained by solving $P_S(\pi, H)$ using Algorithm SP.

**Baseline 2 (Conventional CoMP):** The BS has no cache. The BSs employ CoMP transmission to serve the users by
exchanging both CSI and payload data via backhaul. At each time slot, the precoding matrices are obtained by solving $\mathcal{P}_S (\pi, \mathbf{H})$ using Algorithm SP.

**Baseline 3: (Algorithm SP with Uniform Caching):** The long-term cache control is given by a uniform caching scheme where $q_l = \frac{B_C}{L P_0}$, $\forall l$. The precoding matrices are the same as that of the proposed solution.

We first consider normal user placement. In Fig. [11] we plot the average sum transmit power required for no media playback interruption versus the streaming rate $\mu_0$ for different schemes. The proposed solution has a significant performance gain over baseline 1 (coordinated MIMO), and the gain increases as the BS cache size increases. The proposed solution also has a large gain over baseline 3 (uniform caching) with the same BS cache size. This demonstrates the advantage of the proposed MDS-coded random cache data structure. As the BS cache size increases, the performance of the proposed solution approaches that of baseline 2 (conventional CoMP).

Specifically, when the BS cache size $B_C = 1.8$G Bytes, which is only about half of the total size of all the media files, the performance is already close to baseline 2.

Then we consider edge user placement. In Fig. [12] we plot the average sum transmit power required for no media playback interruption versus the streaming rate $\mu_0$ for different schemes. Similar results as in Fig. [11] can be observed. Moreover, it can be seen that the performance gain of the proposed solution w.r.t. coordinated MIMO is larger under the edge user placement. This shows that a larger CoMP gain can be achieved when the cross links are stronger.

Finally, we show that the proposed solution has significant gain over all the baselines even if we take into account the backhaul overhead. Table [III] compares the average backhaul consumption of different schemes when the streaming rate is 2Mbps and edge user placement is adopted. It can be seen that the average backhaul consumption of the proposed solution is much smaller than the conventional CoMP and the coordinated MIMO. Moreover, the average backhaul consumption decreases as the BS cache size $B_C$ increases. The backhaul consumption under normal user placement is similar and is omitted for conciseness.

**VIII. Conclusion**

We propose a cache-induced opportunistic CoMP scheme for wireless media streaming in MIMO interference networks. By caching a portion of the media files at the BSs, the BSs are able to opportunistically employ CoMP without expensive backhaul. We first propose a novel MDS-coded random cache data structure which can significantly improve the CoMP opportunities. We then formulate a mixed-timescale joint optimization problem for MIMO precoding and cache control. The long-term cache control is used to achieve the best tradeoff between CoMP opportunities and BS cache size. The short-term MIMO precoding is to guarantee the QoS requirements for given cache control. We propose a polynomial complexity precoding algorithm to find a stationary point of the MIMO precoding problem and a stochastic subgradient algorithm to find the cache control solution. The proposed solution can achieve significant performance gain over coordinated MIMO techniques with smaller backhaul loading as demonstrated by numerical simulations.

**APPENDIX**

**A. Proof of Proposition 7**

For convenience, let span ($\mathbf{A}$) represent the subspace spanned by the columns of a matrix $\mathbf{A}$ and orth ($\mathbf{A}$) represent a set of orthogonal basis of span ($\mathbf{A}$). For given precoding matrices $\mathbf{V} = \left\{ \mathbf{V}_{m,k} \in \mathbb{C}^{N_T \times d_{m,k}} : \forall m,k \right\}$, let $\mathbf{V}_{m,k} = \text{orth} \left( \mathbf{H}_{m,k}^{\dagger} \right) \text{orth} \left( \mathbf{H}_{m,k}^{\dagger} \right)^\dagger \mathbf{V}_{m,k}$ denote a precoding matrix obtained by the projection of $\mathbf{V}_{m,k}$ on the subspace span $\left( \mathbf{H}_{m,k}^{\dagger} \right)$. It can be verified that $\mathbf{H}_{m,k} \mathbf{V}_{m,k}' = \mathbf{H}_{m,k} \mathbf{V}_{m,k} \odot \mathbf{H}_{m,k} \mathbf{V}_{m,k}'$ and $\Omega_{m,k} = \Omega_{m,k}' \succeq 0$, where $\Omega_{m,k}$ and $\Omega_{m,k}'$ are the interference-plus-noise covariance matrices respectively resulted from the precoding matrices $\mathbf{V}'$ and $\mathbf{V} = \mathbf{V}_{m,k} \odot \mathbf{H}_{m,k} \mathbf{V}_{m,k}'$. The backhaul consumption of different schemes when the streaming rate is 2Mbps and edge user placement is adopted. It can be seen that the average backhaul consumption of the proposed solution is much smaller than the conventional CoMP and the coordinated MIMO. Moreover, the average backhaul consumption decreases as the BS cache size $B_C$ increases. The backhaul consumption under normal user placement is similar and is omitted for conciseness.

**Figure 11:** Average sum transmit power required for QoS guarantee versus the streaming rate for normal user placement.

**Figure 12:** Average sum transmit power required for QoS guarantee versus the streaming rate for edge user placement.
\( R_k \left( H, V^\prime \right) \geq R_k \left( H, V \right), \quad P_k \left( V^\prime \right) \leq P_k \left( V \right). \quad (22) \)

Since \( \text{Rank} \left( \nabla_{m,k} \right) \leq \text{Rank} \left( H_{m,k} \right) \leq d, \forall m,k \), there exists \( V_{m,k} \in \mathbb{C}^{N_T \times d} \) such that \( V_{m,k} V_{m,k}^\dagger = V_{m,k} V_{m,k}^\dagger, \forall m,k \), from which it follows that

\[
R_k \left( H, V \right) = R_k \left( H, V^\prime \right), \quad P_k \left( V \right) = P_k \left( V^\prime \right), \quad (23)
\]

where \( V = \{ V_{m,k} \in \mathbb{C}^{N_T \times d} : \forall m,k \} \). Then Proposition [1] follows from (22) and (23).

### B. Proof of Proposition [2]

Obviously, we only need to prove that there exists \( \{ V \left( \pi, H \right) : \forall \pi, H \} \) such that \( P \left( V \left( \pi, H \right) \right) < \infty, V \left( \pi, H \right) \in D_\pi \left( \pi, H \right) \) with probability one and \( E \left[ P \left( V \left( \pi, H \right) \right) \right] | \pi \) bounded for all \( \pi \). Consider a simple precoding scheme where \( V_{m,k} \left( \pi, H \right) = 0, \forall k \neq m \) and \( V_{m,k} \left( \pi, H \right) = \sqrt{\frac{1}{2\pi m,k,k}} [e_{m,k,k}, \ldots, e_{m,k,k}] \in \mathbb{C}^{N_T \times d}, \forall k = m \), where \( e_{m,k,k} \) is the dominant eigenvector of \( H_{m,k,k} \) and \( p_k \) is chosen such that

\[
\frac{B_W}{M} \log \left( I + H_{k,k,k} V_{k,k} V_{k,k,k}^\dagger \right) = \frac{1}{2} \ln 2 \mu_{\pi_k} \ln 2. \quad (24)
\]

If \( \sum_{k=1}^K p_k < \infty \), we have \( P \left( V \left( \pi, H \right) \right) \leq \sum_{k=1}^K p_k < \infty \) and \( V \left( \pi, H \right) \in D_\pi \left( \pi, H \right) \). We can verify that

\[
p_k \leq \frac{d}{M} \left[ \text{Tr} \left( H_{k,k,k} H_{k,k,k}^\dagger \right) \right]^{-1} \left( e^{\frac{M \mu_{\pi_k} \ln 2}{M_W} - 1} \right). \quad (24)
\]

Using chain rule, it can be shown that (please refer to [4, Appendix C] for the detailed derivation)

\[
\frac{B_W}{M} \nabla_{V_{m,k}} R_k \left( H, V^*, U^*, W^* \right) = \nabla_{V_{m,k}} R_k \left( H, V^* \right). \quad (27)
\]

It follows from (25), (26) and (27) that

\[
\nabla_{V_{m,k}} L_\pi \left( H, V^* \right) \left( \frac{2}{B_W} \lambda_k \right) = 0, \forall m,k,
\]

where \( L_\pi \left( H, V \right) \) is the Lagrange function of \( P_{S} \left( \pi, H \right) \):

\[
L_\pi \left( H, V \right) = P \left( V \right) + \sum_{k=1}^K \lambda_k \left( \mu_{\pi_k} - R_k \left( H, V^* \right) \right).
\]

Hence, \( V^* \) is a stationary point of \( P_{S} \left( \pi, H \right) \) with \( \frac{2}{B_W} \lambda^* \) as the corresponding Lagrange multiplier vector.

### E. Proof of Lemma [7]

The initial point given in (19) ensures that the initial objective value of \( P_S \left( \pi, H \right) \) is equal to \( P \left( V^* \left( \pi, H \right) \right) \). In each iteration of Algorithm SP, the objective value is strictly decreased before converging to a stationary point of \( P_S \left( \pi, H \right) \). Hence, we must have \( \bar{P} \left( V^* \left( \pi, H \right) \right) \leq P \left( V^* \left( \pi, H \right) \right) \) and

\[
E \left[ \bar{P} \left( V^* \left( \pi, H \right) \right) \right] \leq E \left[ P \left( V^* \left( \pi, H \right) \right) \right] \text{ or } E \left[ P \left( V^* \left( \pi, H \right) \right) \right] \leq \bar{P} \left( V^* \left( \pi, H \right) \right). \text{ (25)}
\]

It can be verified that \( \text{min}_q \{ q_{\pi_k} \} \) is a concave function w.r.t. \( q \). Moreover, it follows from \( E \left[ P \left( V^* \left( \pi, H \right) \right) \right] \leq \bar{P} \left( V^* \left( \pi, H \right) \right) \) that \( \bar{P}_q \left( q, V^* \right) \) is a decreasing linear function of \( \text{min}_q \{ q_{\pi_k} \} \) for \( q \in [0,1] \). Then using the vector composition rule for convex function [17], \( \bar{P}_q \left( q, V^* \right) \) is a convex and non-differentiable function of \( q \) for fixed \( V^* \), which implies that \( P_L \left( V^* \right) \) is also a convex problem.
F. Proof of Lemma 2

Note that $\psi(q, V^*) = E[\phi(q, \pi, H)]$, where

$$\phi(q, \pi, H) = (1 - \min_k \{q_{\pi_k}\}) P(V^*(\pi, H)) + \min_k \{q_{\pi_k}\} \bar{P}(V^*(\pi, H)),$$

is a convex function of $q$. It is easy to see that for each $\pi$ and $H$,

$$G(q, \pi, H) = 1 (l = \pi_k) \left[ \bar{P}(V^*(\pi, H)) - P(V^*(\pi, H)) \right]$$

is a subgradient of $\phi(q, \pi, H)$. Hence, for each $\pi$, $H$ and any $q'$, we have

$$\phi(q', \pi, H) \geq \phi(q, \pi, H) + G(q, \pi, H)^T (q' - q) \quad (28)$$

Taking expectation for both sides of (28) we have

$$\psi(q', V^*) \geq \psi(q, V^*) + E[G(q, \pi, H)]^T (q' - q),$$

which implies that $\bar{G} \triangleq E[G(q, \pi, H)]$ is a subgradient of $\psi(q, V^*)$. On the other hand, we have

$$E \left[ \frac{\partial \psi(q, V^*)}{\partial q_l} \right] = \left[ E \left[ 1 (l = \pi_k) \left( E \left[ \bar{P}(V^*(\pi, H)) \right] \pi \right) \right] - E \left[ P(V^*(\pi, H)) \pi \right] \right] = \bar{G}.$$

Hence, $\frac{\partial \psi(q, V^*)}{\partial q_l}$ is a noisy unbiased subgradient of $\psi(q, V^*)$.

REFERENCES

[1] R. Ghaffar and R. Knopp, “Fractional frequency reuse and interference suppression for OFDMA networks,” in Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2010, pp. 273–277.

[2] O. Somekh, O. Simeone, Y. Bar-Ness, A. Haimovich, and S. Shamai, “Cooperative multicell zero-forcing beamforming in cellular downlink channels,” IEEE Trans. Inf. Theory, vol. 55, no. 7, pp. 3206–3219, 2009.

[3] G. Foschini, K. Karakayali, and R. Valenzuela, “Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency,” IEE Proceedings on Communications, vol. 153, no. 4, pp. 548 – 555, Aug. 2006.

[4] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, “An iteratively weighted minmse approach to distributed sum-utility maximization for a mimo interfering broadcast channel,” IEEE Trans. Signal Processing, vol. 59, no. 9, pp. 4331 –4340, sept. 2011.

[5] A. Liu, Y. Liu, H. Xiang, and W. Luo, “Duality, polite water-filling, and optimization for MIMO B-MAC interference networks and iTree networks,” submitted to IEEE Trans. Info. Theory, Apr. 2010; revised Oct. 2010. [Online]. Available: http://arxiv.org/abs/1004.2484

[6] ——, “MIMO B-MAC interference network optimization under rate constraints by polite water-filling and duality,” IEEE Trans. Signal Processing, vol. 59, no. 1, pp. 263 –276, Jan. 2011.

[7] ——, “Polite water-filling for weighted sum-rate maximization in B-MAC networks under multiple linear constraints,” IEEE Trans. Signal Processing, vol. 60, no. 2, pp. 834 –847, Feb. 2012.

[8] Y. Huang, G. Zheng, M. Bengtsson, K.-K. Wong, L. Yang, and B. Ottersten, “Distributed multicell beamforming with limited intercell coordination,” IEEE Trans. Signal Processing, vol. 59, no. 2, pp. 728–738, 2011.

[9] Y. Huang, C. W. Tan, and B. Rao, “Joint beamforming and power control in coordinated multicell: Max-min duality, effective network and large system transition,” IEEE Trans. Wireless Commun., vol. 12, no. 6, pp. 2730–2742, 2013.

[10] U. Kozat, O. Harmanuc, S. Kanamuri, M. Demircin, and M. Civanlar, “Peer assisted video streaming with supply-demand-based cache optimization,” IEEE Trans. Multimedia, vol. 11, no. 3, pp. 494–508, 2009.

[11] B. Shen, S.-J. Lee, and S. Basu, “Caching strategies in transcoding-enabled proxy systems for streaming media distribution networks,” IEEE Transactions on Multimedia, vol. 6, no. 2, pp. 375–386, 2004.

[12] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, “Femtocaching: Wireless video content delivery through distributed caching helpers,” Proc. IEEE INFOCOM, pp. 1107–1115, 2012.

[13] L. Zheng and D. Tse, “Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels,” IEEE Trans. Info. Theory, vol. 49, no. 5, pp. 1073–1096, 2003.

[14] A. Shokrollahi, “Raptor codes,” IEEE Trans. Info. Theory, vol. 52, no. 6, pp. 2551–2567, 2006.

[15] S. Boyd and L. Xiao, “Subgradient methods,” 2003.

[16] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.

[17] R. S. Sundhar, A. Nedic, and V. V. Veeravalli, “Incremental stochastic subgradient algorithms for convex optimization,” SIAM J. Optim., vol. 20, no. 2, pp. 691–717, 2009.

[18] R. Long Term Evolution of the 3GPP radio technology. 3GPP, 2006. [Online]. Available: http://www.3gpp.org/Highlights/LTE/LTE.htm

[19] Technical Specification Group Radio Access Network; Further Advancements for E-UTRA Physical Layer Aspects. 3GPP TR 36.814. [Online]. Available: http://www.3gpp.org