Coordinated Caching and QoS-Aware Resource Allocation for Spectrum Sharing

Noname manuscript No.
(will be inserted by the editor)

Coordinated Caching and QoS-Aware Resource Allocation for Spectrum Sharing

Konstantinos Ntougias[1] ·
Constantinos B. Papadias[2] ·
Georgios K. Papageorgiou[3] ·
Gerhard Hasslinger[4] ·
Troels B. Sorensen[5]

Received: date / Accepted: date

Abstract 5G cellular networks will heavily rely on the use of techniques that increase the spectral efficiency (SE) to meet the stringent capacity requirements of the envisioned services. To this end, the use of coordinated multi-point (CoMP) as an enabler of underlay spectrum sharing promises substantial SE gains. In this work, we propose novel low-complexity coordinated resource allocation methods based on standard linear precoding schemes that not only maximize the sum-SE and protect the primary users from harmful interference, but they also satisfy the quality-of-service (QoS) demands of the mobile users. Furthermore, we devise coordinated caching strategies that create joint transmission (JT) opportunities, thus overcoming the mobile backhaul / fronthaul throughput and latency constraints associated with the application of this CoMP variant. Additionally, we present a family of caching schemes that outperform significantly the "de facto standard" least recently used (LRU) technique in terms of the achieved cache hit rate while presenting smaller computational complexity. Numerical simulations indicate that the proposed resource allocation methods perform close to their interference-unconstrained counterparts, illustrate that the considered caching strategies facilitate JT, highlight the performance gains of the presented caching schemes over LRU, and shed light on the effect of various parameters on the performance.

Keywords Coordinated QoS-aware interference-constrained power allocation (CQA-ICPA) · Projected zero-forcing (P-ZF) precoding · Cache-aided joint transmission (JT) · Cooperative content caching with redundancy

Konstantinos Ntougias
E-mail: ntougias.konstantinos@ucy.ac.cy
[1] University of Cyprus, Nicosia, Cyprus
[2] American College of Greece, Athens, Greece
[3] Heriot-Watt University, Edinburgh, UK
[4] Deutsche Telekom AG, Darmstadt, Germany
[5] Aalborg University, Aalborg, Denmark

enhancement (C3RE) · Window least-frequently used (WLFU) · Score-gated least-recently used (SG-LRU)

1 Introduction

1.1 Background

The sub-6 GHz spectrum will play a key role in the upcoming 5G cellular networks, as a means to provide the required levels of radio coverage and mobility support [1]. This spectral segment, though, is highly congested [2]. Therefore, the mobile network operators (MNO) will rely on network densification, which enables higher frequency re-use across the service area, as well as on the utilization of techniques that mitigate the inter-cell interference (ICI) and / or increase the spectral efficiency (SE) [1, 2], in order to meet the stringent requirements of the envisioned services in terms of downlink capacity [3].

Coordinated multi-point (CoMP) constitutes an example, wherein neighboring base stations (BS) cooperate with each other to coordinate their resource allocation policies [4,5]. Typically, inter-BS cooperation is limited within clusters to reduce the corresponding overhead. Spectrum sharing is another example, where MNOs access the licensed spectrum of incumbents either by detecting (e.g., via spectrum sensing) and subsequently exploiting idle channels (interweave model) or by maintaining (e.g., via power control) the power of the co-channel interference (CCI) that is received by the incumbents below a predefined threshold (underlay model) [6, 7]. In principle, the latter model presents higher SE gains than the former one (e.g., see [8]).

1.2 Motivation and Related Work

The aforementioned spectrum sharing paradigms have been met with skepticism by the community, because of their inability to provide quality-of-service (QoS) guarantees to the end users. Licensed shared access (LSA) and its enhancements are database-assisted orthogonal spectrum sharing methods that address this issue [9–11]. However, the capacity target of the next-generation cellular networks demands more aggressive re-use of the spectral resources. CoMP can serve as an enabler of underlay spectrum sharing, thus further improving the sum-SE, whilst satisfying the QoS demands of the end users, thanks to its advanced interference management and resource allocation features. Surprisingly enough, though, the relevant works consider the maximization of the sum-SE under a QoS-agnostic context [4,12,13]. Furthermore, these studies consider the use of simple standard linear precoding schemes only in special cases [4].

Moreover, although serving jointly the scheduled users is commonly more efficient than serving disjoint groups of users in a coordinated manner, the latter coordinated beamforming (CBF) variant is applied much more frequently

than the former joint transmission (JT) variant. This is because CoMP-JT, in contrast to CoMP-CBF, requires also the sharing of user data among the cooperating BSs in addition to channel state information (CSI) and control information, thus imposing a heavy burden on the mobile transport network in terms of throughput and latency requirements [4]. Mobile edge caching, wherein servers that have been installed at the network edge (e.g., at the cell sites) store frequently requested content to serve future user requests locally, thus reducing the latency and the network traffic [14–16], provides a workaround to this problem. More specifically, the redundancy of the stored content across the cache servers creates JT opportunities while completely eliminating the need for user data exchanges [17]. Nevertheless, most studies consider uncoordinated transmissions when cache-aided CoMP-JT cannot take place [18, 19]. This approach is highly suboptimal from a sum-SE maximization and interference management perspective and does not suit the underlay spectrum sharing context under consideration.

The caching algorithm that runs on a cache server determines which content will enter or get evicted from the local storage, so that the performance is optimized w.r.t. a given metric. The main performance measure is the cache hit rate, i.e., the fraction of user requests that are served by the cache. Least recently used (LRU) constitutes the most commonly employed caching scheme, due to its simple software implementation, constant $\mathcal{O}(1)$ update effort per request, and ability of adapting to the temporal dynamics of the access pattern. On the other hand, this caching strategy is highly inefficient. Several alternatives that significantly outperform LRU while preserving its beneficial characteristics have been studied in the literature [14, 15, 20], but not under a cache-aided CoMP-JT context.

1.3 Contributions

In this work, we aspire to fill the above-mentioned gaps in the literature. More specifically, we propose QoS-aware and QoS-agnostic coordinated interference-constrained power allocation methods for CoMP-CBF, considering both strict and relaxed interference power thresholds (IPT), as well as a coordinated interference-constrained equal power allocation strategy for CoMP-JT. These techniques are based on the use of standard linear precoding schemes and therefore can be easily adopted by commercial networks. In addition, we devise coordinated caching strategies that create JT opportunities and present simple caching schemes with $\mathcal{O}(1)$ update effort per request that achieve higher cache hit rate than LRU. The performance of the considered resource allocation and caching methods is evaluated via an extensive set of numerical simulations, where CoMP-CBF is utilized when cache-aided CoMP-JT cannot take place.

1.4 Organization and Mathematical Notation

The paper is organized as follows: In Section 2 the system setup is presented, while Section 3 introduces the system model. The proposed resource allocation techniques and their implementation algorithms are given in Section 4, whereas the caching strategies are described in Section 5. The simulation results are discussed in Section 6. Finally, Section 7 provides a summary of this work and presents our conclusions.

Mathematical Notation: $\mathbb{C}$ and $\mathbb{R}$ denote the set of complex and real numbers, respectively, while $\mathbb{R}_+$ denotes the set of non-negative reals. $a \in \mathbb{C}$ is a complex-valued scalar. $\mathbf{a} \in \mathbb{C}^n$ represents a n-dimensional column vector with complex-valued elements. $|a|$ denotes the magnitude (absolute value) of a complex-valued (real-valued) scalar a, whereas $\|\mathbf{a}\|$ stands for the Euclidean norm of $\mathbf{a}$. $\mathbf{A} \in \mathbb{C}^{n \times m}$ represents a $n \times m$ matrix $\mathbf{A}$ with complex-valued entries. $(\mathbf{A})_{*j} = a_{*j}$ denotes the j-th column of $\mathbf{A}$. $\mathbf{A}^{-1}$, $\mathbf{A}^T$, $\mathbf{A}^\dagger$, and $\mathbf{A}^\# := \mathbf{A}^\dagger \left(\mathbf{A}\mathbf{A}^\dagger\right)^{-1}$ denote the inverse, transpose, Hermitian transpose, and Moore-Penrose pseudo-inverse, respectively, of $\mathbf{A}$. $\mathbf{A} = \text{diag}\left(a_1, \ldots, a_n\right)$ denotes a $n \times n$ diagonal matrix $\mathbf{A}$ whose entries on the main diagonal are $(\mathbf{A})_{ii} = a_i$, while $\mathbf{A} = \text{blkdiag}\left(\mathbf{A}_1, \ldots, \mathbf{A}_n\right)$ represents a block diagonal matrix with blocks $\mathbf{A}_i$ on the main diagonal ($i = 1, \ldots, n$). $\mathbf{I}_n$ and $\mathbf{0}_n$ denote the $n \times n$ identity matrix and the n-dimensional null vector, respectively. $\mathbb{E}\left[\cdot\right]$ denotes the expectation operator. $a \sim \mathcal{CN}\left(\mu, \sigma^2\right)$ represents a circularly symmetric complex Gaussian (CSCG) variable a with mean value μ and variance σ^2, while $\mathbf{a} \sim \mathcal{CN}\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n\right)$ represents a CSCG vector $\mathbf{a}$ whose mean value vector and correlation matrix are $\boldsymbol{\mu}$ and $\sigma^2\mathbf{I}_n$, respectively. Finally, $\mathcal{S}$ is a set and $a^+ := \max(0, a)$ for $a \in \mathbb{R}$.

2 System Setup

We consider an underlay spectrum sharing setup wherein the secondary system (SS) is a cellular network and the primary system (PS) is a single-input single-output (SISO) link. The SS is comprised by M cells. In each cell, a BS with N antennas and K active single-antenna mobile stations (MS) are located. Thus, there are M BSs, $N_T = MN$ transmit antennas, and $K_T = MK$ MSs or receive antennas in the cellular network. In CoMP-JT each BS serves all the K_T MSs, whereas in CoMP-CBF it serves only the K MSs of its own cell. Also, each BS is equipped with a cache of storage capacity $C \ll F$ files, where F is the size of the content catalog. We assume files of equal size.

The m-th BS is denoted as BS_m and the k-th MS in the m-th cell is denoted as MS_{km} ($m \in \mathcal{M} = \{1, \ldots, M\}$ and $k \in \mathcal{K} = \{1, \ldots, K\}$). Similarly, the m-th cache server is denoted as C_m. On the other hand, the transmitter and the receiver of the PS are denoted simply as TX_{PS} and RX_{PS}, respectively.

In this work, we assume that: all cells belong to a single cooperation cluster; the BSs serve their users on a single time-frequency resource; the cellular network utilizes universal frequency re-use; the resource allocation decisions

in each cluster (e.g., precoding and power allocation) are taken by a master BS / central unit based on global (cluster-wise) CSI [4]; the mobile transport network is ideal; the nodes have perfect knowledge of the relevant channels; $\mathrm{RX_{PS}}$ informs the master BS about its IPT; and the transmitted symbols and BF vectors are normalized to unit power. Moreover, we consider quasi-static frequency-flat standard i.i.d. Rayleigh fading channels; i.i.d. zero-mean additive white Gaussian noise (AWGN) with unit variance; and i.i.d. Zipf distributed user requests to the files of the catalog [14, 21, 22].

Fig. 1 illustrates the system setup, omitting the cache servers for simplicity. We distinguish between intra-system CCI, which consists of intra-cell multi-user interference (MUI) and ICI components, and forward / reverse inter-system (FIS / RIS) CCI, as shown in this figure.

3 System Model

The channel between MS_{km} and BS_j is denoted as $\mathbf{h}_{km}^{j} \sim \mathcal{CN}\left(\mathbf{0}_N, \mathbf{I}_N\right)$, while the BF vector, transmission power, and transmitted data symbol associated with this link are denoted as $\mathbf{w}_{mk}^{j} \in \mathbb{C}^N$, $P_{mk}^{j} \in \mathbb{R}_+$, and $s_{mk}^{j} \sim \mathcal{CN}(0,1)$, respectively ($k \in \mathcal{K};\ m,\ j \in \mathcal{M}$). Similarly, the channel between $\mathrm{RX_{PS}}$ and $\mathrm{TX_{PS}}$ is denoted as $g \sim \mathcal{CN}(0,1)$, while the transmission power and transmitted data symbol associated with this link are denoted as $P \in \mathbb{R}_+$ and $d \sim \mathcal{CN}(0,1)$, respectively. On the other hand, the interfering channels between MS_{km} and $\mathrm{TX_{PS}}$ and between $\mathrm{RX_{PS}}$ and BS_j are denoted as $h_{km} \sim \mathcal{CN}(0,1)$ and $\mathbf{g}^{j} \sim \mathcal{CN}\left(\mathbf{0}_N, \mathbf{I}_N\right)$, respectively. Finally, the AWGN at MS_{km} and $\mathrm{RX_{PS}}$ is denoted as $n_{km} \sim \mathcal{CN}(0,1)$ and $z \sim \mathcal{CN}(0,1)$, respectively.

3.1 Coordinated Beamforming

System Model Assuming the application of CoMP-CBF, the complex baseband representation of the received signal at MS_{km} at a given time sample, $y_{km} \in \mathbb{C}$, is:

$$y_{km} = \sum_{j=1}^{M} \sum_{l=1}^{K} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{jl}^{j} + h_{km}\sqrt{P}d + n_{km}, \tag{1}$$

where $\mathbf{v}_{jl}^{j} \in \mathbb{C}^N$ is expressed as:

$$\mathbf{v}_{jl}^{j} = \mathbf{w}_{jl}^{j}\sqrt{P_{jl}^{j}}s_{jl}^{j}, \quad j \in \mathcal{M},\ l \in \mathcal{K}. \tag{2}$$

In Eq. (1) we have omitted the time index, for convenience. Also, we have implicitly assumed that $k \in \mathcal{K}$ and $m \in \mathcal{M}$. We follow these practices throughout the manuscript. The first term at the right-hand-side (RHS) of Eq. (1) can be

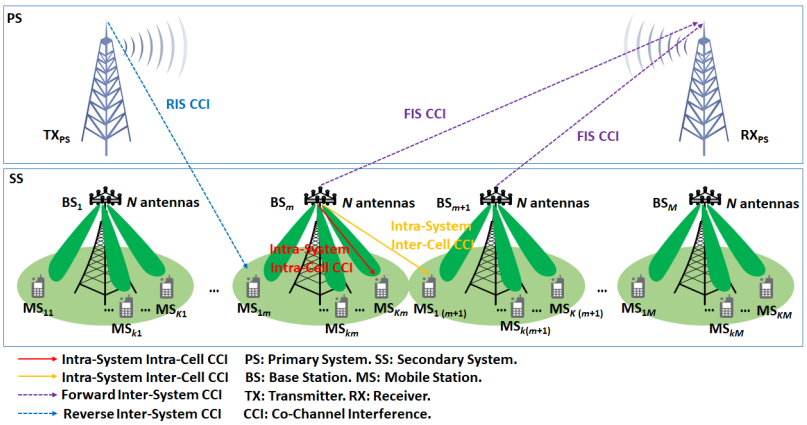


Fig. 1: System setup, notation, and types of interference.

decomposed into the sum of a data component, an intra-cell MUI component, and an ICI component as follows:

$$\breve{y}_{km} = \left(\mathbf{h}_{km}^{m}\right)^{\dagger} \mathbf{v}_{mk}^{m} + \sum_{\substack{i=1 \\ i \neq k}}^{K} \left(\mathbf{h}_{km}^{m}\right)^{\dagger} \mathbf{v}_{mi}^{m} + \sum_{\substack{j=1 \\ j \neq m}}^{M} \sum_{l=1}^{K} \left(\mathbf{h}_{km}^{j}\right)^{\dagger} \mathbf{v}_{jl}^{j}. \tag{3}$$

The remaining terms at the RHS of Eq. (1) represent, from left to right, the RIS CCI and the AWGN at MS_{km}.

The complex baseband representation of the received signal at $\mathrm{RX}_{\mathrm{PS}}$, $y \in \mathbb{C}$, is given by:

$$y = g\sqrt{P}d + \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\mathbf{g}^{m}\right)^{\dagger} \mathbf{v}_{mk}^{m} + z. \tag{4}$$

The first term at the RHS of Eq. (4) is the data component, the next one is the FIS CCI component, and the last one is the AWGN at $\mathrm{RX}_{\mathrm{PS}}$.

SINR, Data Rate, and Sum-Rate The signal-to-interference-plus-noise-ratio (SINR) of MS_{km}, $\gamma_{km} \in \mathbb{R}$, is given by:

$$\gamma_{km} = \frac{\left|\mathbf{v}_{mk}^{m}\right|^{2}}{\sum\limits_{\substack{i=1 \\ i \neq k}}^{K} \left|\mathbf{v}_{mi}^{m}\right|^{2} + \sum\limits_{\substack{j=1 \\ j \neq m}}^{M} \sum\limits_{l=1}^{K} \left|\mathbf{v}_{jl}^{j}\right|^{2} + I_{km}}, \tag{5}$$

where

$$\left|\mathbf{v}_{mk}^{m}\right|^{2} = \left|\left(\mathbf{h}_{km}^{m}\right)^{\dagger} \mathbf{w}_{mk}^{m}\right|^{2} P_{mk}^{m} \tag{6}$$

and

$$I_{km} = \left|h_{km}\right|^{2} P + 1. \tag{7}$$

The nominator in Eq. (5) corresponds to the power of the data signal component that is received at MS_{km}. I_{km} is the sum of the powers of the RIS CCI and the AWGN at MS_{km}. The remaining terms at the denominator of Eq. (5) represent, from left to right, the power of the intra-cell MUI and ICI components that are received at MS_{km}.

If we assume that the transmitted symbols are drawn by a zero-mean complex Gaussian distribution, then the data rate of this user, $R_{km} \in \mathbb{R}_+$, is given by the Shannon formula:

$$R_{km} = \log_2 (1 + \gamma_{km}) . \tag{8}$$

Finally, the sum-rate (SR) of the cellular network per unit of spectral bandwidth (i.e., the sum-SE), $R \in \mathbb{R}_+$, is given by:

$$R = \sum_{m=1}^{M} \sum_{k=1}^{K} R_{km}. \tag{9}$$

Composite Block Matrix Representation The composite system model of the SS in block matrix form, ignoring the RIS CCI for convenience, is given by:

$$\mathbf{y}_{\mathrm{SS}} = \mathbf{H}\mathbf{W} \left(\mathbf{P}_{\mathrm{SS}}\right)^{1/2} \mathbf{s} + \mathbf{n}. \tag{10}$$

where $\mathbf{y}_{\mathrm{SS}} = \begin{bmatrix} y_{11} \cdots y_{KM} \end{bmatrix}^T$, $\mathbf{s} = \begin{bmatrix} s_{11} \cdots s_{MK} \end{bmatrix}^T$, and $\mathbf{n} = \begin{bmatrix} n_{11} \cdots n_{KM} \end{bmatrix}^T$ are the received symbols vector, transmitted symbols vector, and AWGN vector, respectively.

The composite channel matrix in Eq. (10) is:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \vdots \\ \mathbf{H}_M \end{bmatrix}, \tag{11}$$

where

$$\mathbf{H}_m = \begin{bmatrix} \left(\mathbf{h}_{1m}^1\right)^\dagger & \cdots & \left(\mathbf{h}_{1m}^M\right)^\dagger \\ \vdots & \ddots & \vdots \\ \left(\mathbf{h}_{Km}^1\right)^\dagger & \cdots & \left(\mathbf{h}_{Km}^M\right)^\dagger \end{bmatrix}. \tag{12}$$

That is, $\mathbf{H}_m$ holds the channels of the K users in the m-th cell with the M BSs. More specifically, the k-th row holds the channels of the k-th user in the m-th cell MS_{km} with $\mathrm{BS}_1, \ldots, \mathrm{BS}_M$.

The precoding matrix in Eq. (10) is:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \cdots \mathbf{W}_M \end{bmatrix}, \tag{13}$$

where

$$\mathbf{W}_m = \begin{bmatrix} \mathbf{w}_{m1}^1 & \cdots & \mathbf{w}_{mK}^1 \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{m1}^M & \cdots & \mathbf{w}_{mK}^M \end{bmatrix}. \tag{14}$$

That is, $\mathbf{W}_m$ holds the BF vectors of the M BSs for the users in the m-th cell. More specifically, the k-th column holds the BF vectors of $\mathrm{BS}_1, \ldots, \mathrm{BS}_M$ for the k-th user in the m-th cell MS_{km}.

The power allocation (PA) matrix in Eq. (10) is:

$$\mathbf{P}_{\mathrm{SS}} = \mathrm{blkdiag}\left(\mathbf{P}_1, \ldots, \mathbf{P}_M\right), \tag{15}$$

where

$$\mathbf{P}_m = \mathrm{diag}\left(P_{m1}^m, \ldots, P_{mK}^m\right). \tag{16}$$

That is, $\mathbf{P}_m$ holds the powers allocated by BS_m to its K users $\mathrm{MS}_{1m}, \ldots, \mathrm{MS}_{Km}$.

3.2 Joint Transmission

Similar to the CoMP-CBF case, the complex baseband representation of the received signal at MS_{km} when CoMP-JT is applied is given by:

$$y_{km} = \sum_{j=1}^{M} \sum_{l=1}^{M} \sum_{i=1}^{K} \left(\mathbf{h}_{km}^j\right)^{\dagger} \mathbf{v}_{li}^j + h_{km}\sqrt{P}d + n_{km}. \tag{17}$$

The first term at the RHS of Eq. (17) can be decomposed into the sum of a data component, an intra-cell MUI component, and an ICI component as follows:

$$\breve{y}_{km} = \sum_{j=1}^{M} \left(\mathbf{h}_{km}^j\right)^{\dagger} \mathbf{v}_{mk}^j + \sum_{j=1}^{M} \sum_{\substack{i=1 \\ i \neq k}}^{K} \left(\mathbf{h}_{km}^j\right)^{\dagger} \mathbf{v}_{mi}^j + \sum_{j=1}^{M} \sum_{\substack{l=1 \\ l \neq m}}^{M} \sum_{i=1}^{K} \left(\mathbf{h}_{km}^j\right)^{\dagger} \mathbf{v}_{li}^j. \tag{18}$$

The complex baseband representation of the received signal at $\mathrm{RX}_{\mathrm{PS}}$ is given by:

$$y = g\sqrt{P}d + \sum_{m=1}^{M} \sum_{l=1}^{M} \sum_{k=1}^{K} \left(\mathbf{g}^m\right)^{\dagger} \mathbf{v}_{lk}^m + z. \tag{19}$$

The SINR expression and composite block matrix system model for the CoMP-JT case can be easily derived by these equations and are omitted here. In the following sections, we will assume that CoMP-CBF is applied, unless it is explicitly stated otherwise.

4 Resource Allocation Methods and Algorithms

4.1 Transmission Constraints

Each transmission from a BS to one of its K users has non-negative power:

$$P_{mk}^m \geq 0. \tag{20}$$

Furthermore, the transmissions within the SS are subject to a sum-power constraint (SPC) per BS, i.e., each BS is allowed to serve its K users with a total power that does not exceed a maximum value P_T:

$$\sum_{k=1}^{K} P_{mk}^{m} \leq P_T. \tag{21}$$

In addition, the operation of the SS is subject to an interference power constraint (IPC), i.e., the total power of the FIS CCI component that is received at $\mathrm{RX_{PS}}$ should not exceed an IPT P_I. By defining:

$$\alpha_{mk}^{m} = \left| (\mathbf{g}_m)^{\dagger} \mathbf{w}_{mk}^{m} \right|^2, \tag{22}$$

we can express the IPC as follows:

$$\sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{mk}^{m} P_{mk}^{m} \leq P_I. \tag{23}$$

In many cases, we should ensure that the data rate of MS_{km} is at least equal to a minimum value $\tilde{R}_{km} > 0$. This QoS constraint is expressed as:

$$R_{km} \geq \tilde{R}_{km}, \tag{24}$$

or, in view of Eq. (8), w.r.t. the minimum required SINR $\tilde{\gamma}_{km}$:

$$\gamma_{km} \geq \tilde{\gamma}_{km}. \tag{25}$$

4.2 General Optimization Problems

Problem **P1.A** refers to the joint determination of the BF vectors and allocated powers that maximize the SR under the constraints of Section 4.1:

$$\max_{\substack{\mathbf{w}_{mk}^{m},\ P_{mk}^{m} \\ m\in\mathcal{M},\ k\in\mathcal{K}}} \quad R = \sum_{m=1}^{M} \sum_{k=1}^{K} \log_2 \left(1 + \lambda_{mk}^{m} P_{mk}^{m}\right) \tag{26a}$$

s.t.

$$\sum_{k=1}^{K} P_{mk}^{m} \leq P_T, \quad m \in \mathcal{M}, \tag{26b}$$

$$\sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{mk}^{m} P_{mk}^{m} \leq P_I, \tag{26c}$$

$$P_{mk}^{m} \geq \tilde{P}_{mk}^{m}, \quad k \in \mathcal{K},\ m \in \mathcal{M}, \tag{26d}$$

where $\lambda_{mk}^{m} = \gamma_{km}/P_{mk}^{m}$ and the per-user QoS constraints in Eq. (26d) are derived from Eq. (25) by substituting $\gamma_{km} = \lambda_{mk}^{m} P_{mk}^{m}$ and $\tilde{\gamma}_{km} = \lambda_{mk}^{m} \tilde{P}_{mk}^{m}$,

with $\tilde{P}^m_{mk}$ denoting the minimum power that should be allocated to MS_{km}. By setting $\tilde{P}^m_{mk} = 0$, the per-user QoS constraints in Eq. (26d) are converted into the non-negative allocated power per-user constraints of Eq. (20) and the resulting optimization problem is referred to as **P1.B**. If, in addition, we omit the IPC of Eq. (26c), we obtain **P1.C**. Note that these optimization problems are non-convex, due to the inter-user coupling through the interference components at the denominator of the SINR term in the expression of the SR.

4.3 Relaxed Optimization Problems based on Zero-Forcing Precoding

Let us assume the application of coordinated zero-forcing (ZF) precoding at the SS in a spectrum-sharing-agnostic manner, i.e., by ignoring the inter-system CCI terms. The ZF precoding matrix $\mathbf{W}^{(\mathrm{ZF})}$ is obtained as follows [4]:

$$\mathbf{F}^{(\mathrm{ZF})} = \mathbf{H}^{\#} = \mathbf{H}^{\dagger} \left(\mathbf{H}\mathbf{H}^{\dagger}\right)^{-1}. \tag{27a}$$

$$\mathbf{W}^{(\mathrm{ZF})} = \frac{\left(\mathbf{F}^{(\mathrm{ZF})}\right)_{*j}}{\left\| \left(\mathbf{F}^{(\mathrm{ZF})}\right)_{*j} \right\|}, \quad j = 1, \ldots, K_T. \tag{27b}$$

ZF eliminates the intra-system CCI within the cellular network (i.e., the intra-cell MUI and the ICI). That is,

$$\sum_{\substack{i=1 \\ i \neq k}}^{K} \left(\mathbf{h}^m_{km}\right)^{\dagger} \left(\mathbf{v}^m_{mi}\right)^{(\mathrm{ZF})} + \sum_{\substack{j=1 \\ j \neq m}}^{M} \sum_{i=1}^{K} \left(\mathbf{h}^j_{km}\right)^{\dagger} \left(\mathbf{v}^j_{ji}\right)^{(\mathrm{ZF})} = 0. \tag{28}$$

Thus, the SINR of MS_{km} becomes:

$$\gamma_{km} = \frac{\left| \left(\mathbf{h}^m_{km}\right)^{\dagger} \left(\mathbf{w}^m_{mk}\right)^{(\mathrm{ZF})} \right|^2 P^m_{mk}}{\left|h_{km}\right|^2 P + 1}. \tag{29}$$

After the application of ZF precoding, **P1.A**-**P1.C** are converted into the corresponding convex PA tasks **P2.A**-**P2.C**, since the coupled interference terms have been removed from the objective function.

4.4 Solutions to the Relaxed Optimization Problems and Algorithms

The solutions to the aforementioned PA tasks are presented in the following theorem [23].

Theorem 1 *The solution to **P2.A** is given by the coordinated QoS-aware interference-constrained PA (CQA-ICPA) scheme:*

$$P^m_{mk} = \left(\frac{1}{\ln 2 \left(\nu_m + \mu \alpha^m_{mk}\right)} - \frac{1}{\lambda^m_{mk}} - \tilde{P}^m_{mk} \right)^{+} + \tilde{P}^m_{mk}, \tag{30}$$

where ν_m and μ are Lagrange multipliers associated with the SPCs and the IPC, respectively.

*Similarly, the solution to **P2.B** is the coordinated ICPA (C-ICPA) scheme that is obtained from Eq. (30) by setting $\tilde{P}_{mk}^m = 0$:*

$$P_{mk}^m = \left(\frac{1}{\ln 2 \left(\nu_m + \mu \alpha_{mk}^m \right)} - \frac{1}{\lambda_{km}} \right)^+ . \tag{31}$$

These solutions are obtained by taking the Lagrangian form of the corresponding optimization problems and applying the Karush-Kuhn-Tucker (KKT) conditions [24]. The iterative algorithm that calculates the Lagrange multipliers and implements these solutions is presented in Algorithm 1. The algorithm, whose accuracy is controlled by the parameter $\delta_\mu > 0$, makes use of the bisection method to update the value of μ in each iteration based on whether the IPC is met or not. Note that when the IPC is inactive, $\mu = 0$ and the algorithm reduces to the corresponding interference-unconstrained PA (IUPA) solution, that is, to CQA-IUPA or to C-IUPA. The latter is the solution to **P2.C** and corresponds to the standard coordinated water-filling (WF) PA scheme for a stand-alone CoMP-CBF cellular setup [4].

4.5 Solutions for Alternative Linear Precoding Schemes

Maximum ratio transmission (MRT) is an egoistic precoding strategy that matches the BF vector of each user to its channel vector, in order to maximize the receive signal-to-noise-ratio (SNR) [4]:

$$(\tilde{\mathbf{w}}_{mk}^m)^{(\mathrm{MRT})} = \mathbf{h}_{km}^m ; \ (\mathbf{w}_{mk}^m)^{(\mathrm{MRT})} = \frac{(\tilde{\mathbf{w}}_{mk}^m)^{(\mathrm{MRT})}}{\left\| (\tilde{\mathbf{w}}_{mk}^m)^{(\mathrm{MRT})} \right\|}. \tag{32}$$

In contrast to the altrouistic ZF precoding scheme, MRT performs well in the noise-limited low-SNR regime, since it focuses the radiated power towards the intended users, but its capacity floors in the interference-limited high-SNR regime, due to the uncoordinated CCI.

Regularized ZF (RZF), on the other hand, is an extension of ZF precoding that maximizes the SINR at each user, in order to improve the performance in the low-SNR regime. That is, we have for $j = 1, \ldots, K_T$ [4]:

$$\mathbf{F}^{(\mathrm{RZF})} = \mathbf{H}^\dagger \left(\frac{1}{K_T} \mathbf{I}_{K_T} + \mathbf{H}\mathbf{H}^\dagger \right)^{-1} ; \ \mathbf{W}^{(\mathrm{RZF})} = \frac{\left(\mathbf{F}^{(\mathrm{RZF})} \right)_{*j}}{\left\| \left(\mathbf{F}^{(\mathrm{RZF})} \right)_{*j} \right\|}. \tag{33}$$

Although these precoding techniques do not eliminate the intra-system CCI within the SS, we can apply heuristically the PA solutions derived in Section 4.4 for ZF precoding [4].

Algorithm 1 CQA-ICPA and C-ICPA Algorithm for CBF

procedure CQA-ICPA($\lambda_{mk}^m, \alpha_{mk}^m, P_T, P_I, \tilde{P}_{mk}^m$)
  Initialize: $\mu_{\min}$, $\mu_{\max}$
  while $|\mu_{\max} - \mu_{\min}| > \delta_\mu$ **do**
    $\mu = (\mu_{\min} + \mu_{\max})/2$
    if (ICPA) then
      Set: $\tilde{P}_{mk}^m = 0$
    end if
    for $m = 1$ to M **do**
      Find $\min(\nu_m)$, $\nu_m \geq 0$:
      $\sum_{k=1}^{K} \left(\left(\frac{1}{\ln 2 (\nu_m + \mu \alpha_{mk}^m)} - \frac{1}{\lambda_{mk}^m} - \tilde{P}_{mk}^m \right)^+ + \tilde{P}_{mk}^m \right) \leq P_T$
    end for
    Compute P_{mk}^m according to Eq. (30) (CQA-ICPA) or Eq. (31) (ICPA)
    if $\sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{mk}^m P_{mk}^m \geq P_I$ **then**
      $\mu_{\min} = \mu$
    else
      $\mu_{\max} = \mu$
    end if
  end while
  Output: $P_{mk}^m, \quad m \in \mathcal{M};\ k \in \mathcal{K}$
end procedure

4.6 Suboptimal Power Allocation Methods

A simple suboptimal PA method for both the CoMP-CBF and CoMP-JT cases would be to allocate equal power to the users, taking though into account the SPCs and the IPC, as shown in Proposition 1.

Proposition 1 *The coordinated interference-constrained equal PA (C-ICEPA) scheme allocates the following power to each user:*

$$P_c = \begin{cases} \min\left(\frac{P_T}{K}, \frac{P_I}{\sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{mk}^m} \right) & (\textit{CoMP-CBF}) \\ \min\left(\frac{P_T}{MK}, \frac{P_I}{\sum_{j=1}^{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{mk}^j} \right) & (\textit{CoMP-JT}) \end{cases}. \tag{34}$$

Note that when the IPC is inactive, C-ICEPA is reduced to conventional C-IUEPA, i.e., $P_c = P_T/K$ for CoMP-CBF or $P_c = P_T/MK$ for CoMP-JT.

4.7 Projected ZF Precoding

The composite FIS CCI channel $\mathbf{g} \in \mathbb{C}^{N_T}$ is defined as:

$$\mathbf{g} = \left[\mathbf{g}^1 \cdots \mathbf{g}^M \right]. \tag{35}$$

Let us define:

$$\hat{\mathbf{g}} = \frac{\mathbf{g}^\dagger}{\|\mathbf{g}\|}. \tag{36}$$

The projection of the composite channel matrix $\mathbf{H}$ into the null space of $\hat{\mathbf{g}}$ is given by:

$$\hat{\mathbf{H}} = \mathbf{H}\left(\mathbf{I}_{N_T} - \hat{\mathbf{g}}\left(\hat{\mathbf{g}}\right)^{\dagger}\right). \tag{37}$$

When the IPT is hard, we can compute the ZF precoders based on $\hat{\mathbf{H}}$ and then apply conventional C-IUPA or CQA-IUPA, instead of computing them based on $\mathbf{H}$ and applying C-ICPA or CQA-ICPA, respectively. This method is called projected ZF (P-ZF) precoding with C-IUPA / CQA-IUPA and ensures that the FIS CCI is completely eliminated.

5 Coordinated Caching

5.1 Zipf's Law

Several measurements of the patterns of requests for content on the Internet revealed that the user behavior is governed by Zipf's law, so that a small subset of popular objects (e.g., videos, files, web site pages, etc.) attracts the main portion of the user requests. In particular, a Zipf distribution associated with a finite set $\mathcal{F}$ of F objects O_r attributes request probabilities $p(r)$ to these objects corresponding to their popularity rank $r = 1, \ldots, F$, with $p(1) \geq \cdots \geq p(F)$, according to the following relation:

$$p(r) = Ar^{-\beta} = \frac{r^{-\beta}}{\sum_{r=1}^{F} r^{-\beta}}, \quad A, \beta > 0, \tag{38a}$$

$$\sum_{r=1}^{F} p(r) = 1 \Rightarrow A = p(1) = \frac{1}{\sum_{r=1}^{F} r^{-\beta}}, \tag{38b}$$

where A is a normalization constant and β is a shaping factor that determines the skewness of the distribution. Typically, $\beta \in \{0.5, 1\}$ [21, 22, 25, 26].

The high concentration of user requests to a small number of popular objects implies that even relatively small caches can be quite efficient, in terms of the achieved cache hit rate, provided that the applied caching strategy stores the most popular objects in the cache.

5.2 Content Popularity Dynamics

In Section 5.1, we have implicitly assumed a stream of i.i.d. requests to the objects, so that a request refers to an object O_r with a constant probability $p(r)$. Under this independent reference model (IRM) [27], the optimum hit rate of a cache with storage capacity sufficient to hold $C \ll F$ objects equals the sum of the access probabilities of the top C objects in terms of popularity ranking, i.e., $h_{\text{opt}}^{C} = \sum_{r=1}^{C} p(r)$.

In practice, though, the popularity of the objects changes over time and new objects enter the content catalog. Nevertheless, the popularity dynamics is

in general relatively low (i.e., rank changes take place in the time scale of hours or days and affect only a small subset of the objects each time [21, 26]). For such slowly varying pattern, the hit rates achieved by caches that serve a large user population and handle hundreds of thousands of requests per day are close to the ones computed under IRM conditions for Zipf distributed requests [20]. Thus, Eq. (38) represents in most cases a simple, yet valid approximation of content access patterns.

Therefore, we can approach the optimum hit rate in practice if we hold in the cache the most popular objects over long timeframes [14, 15, 20]. On the other hand, a practical caching scheme should be able to react to the popularity changes observed in realistic scenarios by replacing formerly popular cached objects with new ones that became "hot" in a recent timeframe, in order to avoid the pollution of the cache storage by outdated objects.

5.3 Caching Schemes

A caching scheme assigns values to the objects either explicitly according to a score function or implicitly via a ranking method, in order to determine which objects to store in or drop from the local storage. It is typically implemented in software as some type of list of stored objects. The computational complexity of the cache storage lookup, cache update, and object insertion / replacement operations is of major importance in practical implementations. Several caching schemes whose goal is to maximize the cache hit rate have been studied in the literature.

LRU and LFU Least recently used (LRU) ranks the objects according to their time-of-last-access and stores in a cache of size C the C most recently referenced objects, sorted in decreasing request recency order. LRU is the most widely adopted caching scheme due to its simple implementation, constant $\mathcal{O}(1)$ effort per request for putting the requested object on the top of the cache stack, and ability to adapt to the access pattern dynamics, since it promotes the caching of recently "active" objects [14]. On the other hand, this caching scheme is highly inefficient, as it has been shown analytically as well as through numerical simulations and trace-based measurement studies, because it does not take into account object popularity in the caching and replacement decisions. In fact, the absence of request count statistics leads often to the pollution of LRU caches by objects that are referenced only once, which degrades the caching efficiency [14,15,20]. Moreover, LRU presents a high rate of loading objects into the cache, since in each cache miss the requested object is transferred to the cache storage. The frequent downloading of external objects increases the processing load, the latency, and the network traffic.

Least frequently used (LFU), on the other hand, counts the number of past requests to each object and holds in a cache of size C the C most frequently referenced objects, sorted in decreasing request frequency order. Typically, LRU is used as a tie-breaker between objects of the same value. LFU achieves

the optimum hit rate under IRM conditions, since its request count statistics converge over time and reflect the popularity ranking of the objects. Also, LFU allows the caching of requested objects only when their request count is higher than (or at least equal to) the request count of the least frequently referenced cached object, thus reducing the loading rate of external objects into the cache. However, this caching strategy is rarely used in practical applications, since its unlimited request statistics leads to pollution of the cache by currently irrelevant objects that are maintained in the local storage over long timeframes due to their high request count and influece the caching and replacement decisions. Yet, LFU serves as a benchmark under the IRM. Another reason why LFU has not been preferred in practice is because the conventional implementation of this caching scheme presents an $\mathcal{O}(C)$ insertion, replacement, and update complexity for maintaining a perfectly sorted (w.r.t. the request count of the objects) cache list of size C. Nevertheless, we should note that there have been proposed also implementations of LFU with $\mathcal{O}(1)$ effort per request (which, however, require at least twice the time needed by LRU to perform a cache update) [28].

Design Criteria Several alternatives to these standard caching methods have been proposed in the literature. Our focus is on caching schemes that meet the following criteria [20]:

1. They have simple implementation and present constant $\mathcal{O}(1)$ effort per request.
2. They approach the optimum LFU hit rate under IRM conditions.
3. They react to the dynamically changing popularity of the objects.
4. They implement some admission control and replacement policy that reduces the rate of loading external objects into the cache.
5. They provide the flexibility to consider other performance metrics than the cache hit rate.

Typically, such caching strategies inspect access statistics of past requests to extract information about the frequency and recency of requests to objects.

WLFU and WLFU-NE Window LFU (WLFU) [14] restricts the LFU principle in a sliding window (SW) of W requests, which acts as an aging mechanism that prevents cache pollution with objects of decreasing relevance. The window size determines the reach of the statistics in the past and, thus, represents a single adaptation parameter for balancing the impact of request frequency and recency information on caching and replacement decisions. WLFU resembles LRU for small window sizes and approaches LFU as the window size increases. We can further simplify this caching method by performing insertion of objects always from the beginning of the cache list (i.e., in an LRU-like fashion) and by considering simple cache updates that involve the comparison of the scores of the requested object upon a cache hit or of a cached object whose request dropped from the window and its neighbor from left or from right in the cache list, respectively. This WLFU with neighbor (position) exchange (WLFU-NE)

cache updates scheme starts with a cache list that is partially sorted w.r.t. the objects' scores and over time results in a perfectly sorted cache list via the aforementioned simple cache updates [15].

SG-LRU and SG-C By using LRU-type updates instead of WLFU-NE updates, we get score-gated LRU (SG-LRU) [20]: a caching scheme that combines the LRU principle for simple implementation and fast updates with a score-gate function (here, WLFU) for avoiding the frequent loading of objects in the cache and storing the most popular objects in a recent timeframe. Fig. 2 depicts the operation of this caching scheme with the help of an example. SG-LRU runs faster than LRU, since it avoids the frequent updates caused by cache misses. Moreover, it replaces over time lower ranked objects with higher ranked ones in the cache and approximates closely WLFU / WLFU-NE. If we omit the LRU cache structure (i.e., if upon a cache hit we simply update the score of the requested object but not its position in the cache list) and compare in each user request the score of the requested object with the score of a random cached object that is determined by a corresponding pointer ("clock hand") that cycles through the cache list, then we will obtain an even simpler score-gated clock (SG-C) scheme [28]. SG-C runs faster than SG-LRU due to the fact that the LRU updates caused by cache hits are relatively time-consuming operations, in contrast to the LRU updates caused by cache misses. Notice that SG-LRU and SG-C provide the flexibility to use an arbitrary scoring function for ranking the objects and, therefore, can optimize the performance w.r.t. any criterion. Thus, these caching strategies meet all the design criteria mentioned previously. This is in contrast to WLFU and WLFU-NE, where the scoring function defines also and the cache structure / caching principle. Hence, these caching schemes do not meet the design criterion (5).

5.4 C3RE Caching

In this work, we propose a coordinated content caching with redundancy enhancement (C3RE) method, where upon a local cache miss the target cache downloads the requested object from another cache in the cluster if possible (global cache hit), thus leaving the fetching of this object from the origin server as a last resort (global cache miss). Upon a global cache hit, the remote cache may update only its window (cooperation variant II.A) or both its window and local storage (cooperation variant II.B) or none of them (cooperation variant I), assuming that WLFU, WLFU-NE, or SG-LRU is utilized. When SG-C is employed, only the cooperation variants I and II.A (simply called variant II in this case) are valid. Similarly, when LRU is applied, the remote cache may (variant II) or may not (variant I) update its local storage upon a global cache hit. On the other hand, whenever a file enters the target cache, the corresponding BS updates both the window and local storage of its cache, unless SG-C which involves only score updates is used.

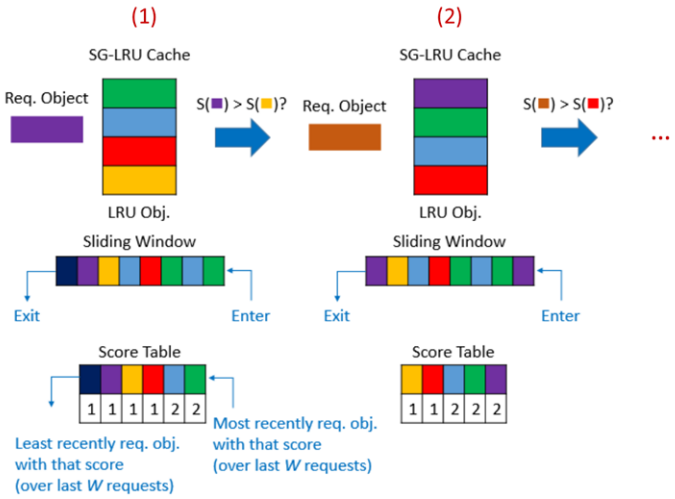


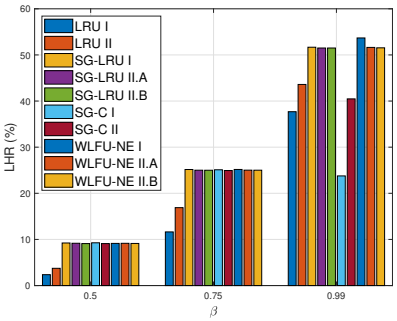
Fig. 2: Replacement operation of an SG-LRU cache with size $C = 4$ that utilizes the WLFU score function with a sliding window of size $W = 8$.

Cooperative caching reduces the latency and network traffic associated with the downloading of objects from the origin server upon local cache misses at the expense of some cooperation overhead to fetch the requested object from a remote cache. Furthermore, the aforementioned caching variants create different levels of content redundancy across the cache servers, which can be exploited towards JT transmissions.

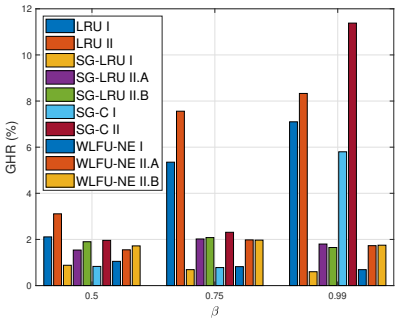
6 Performance Evaluation via Numerical Simulations

6.1 Cache Hit Rates

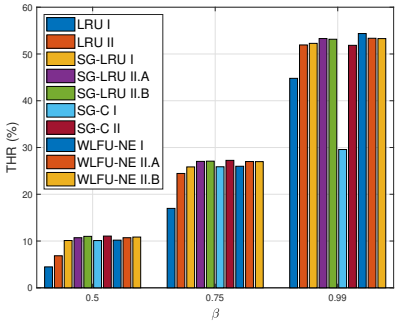
In this section, we study the performance of the considered C3RE variants when LRU, SG-LRU, SG-C, and WLFU-NE is applied, in terms of the average local, global, and total cache hit rate (LHR, GHR, and THR, respectively) that is achieved after $N_{\text{sim}} = 1,000$ simulation runs. We assume a cooperation cluster consisting of $M = 2$ cells, with $K = 1$ active user per cell that has been selected from a large user population via some user scheduling algorithm. We also assume initially a content catalog with a size of $F = 10,000$ files, $N_c = 2$ cache servers (one for each BS) with a storage capacity of $C = 100$ files each (i.e., equal to the 1% of the catalog size), $N_r = 1,000,000$ user requests addressed to each cache server in every simulation run, and a window with a size of $W = 100,000$ requests (i.e., equal to the 10% of the requests). Note that in each simulation run, we ignore the results of the first 25% of the requests, to exclude the cache filling phase and transient behavior from the steady-state performance evaluation.



(a) Local hit rates.



(b) Global hit rates.



(c) Total hit rates.

Fig. 3: Cache hit rates for varying Zipf shape parameter β.

Use Case (1): We vary the shape parameter of the Zipf distribution as $\beta = \{0.5; 0.75; 0.99\}$. The hit rates are shown in Fig. 3.

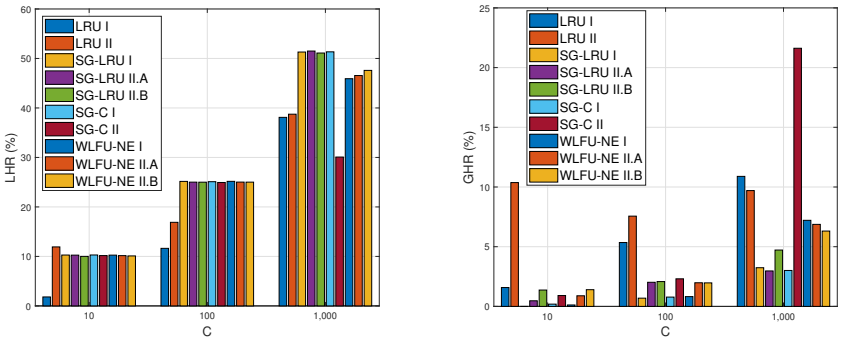
– LHR: *The LHR of all caching strategies improves as β increases*, as expected. For LRU, this is attributed to the fact that as the popular objects become "hotter", they are typically requested more often in shorter time intervals and, therefore, enter more frequently the LRU cache. The statistics-based caching schemes achieve similar LHR across the considered range of β, except for the SG-C variants whose LHR degrades for high β due to the random selection of the "least valuable" cached object, which does not let this strategy to exploit the high unbalance of user requests in favor of popular objects in this scenario. Naturally, *the statistics-based caching schemes outperform significantly LRU*, due to the exploitation of request count information in the caching and replacement decisions. We should also mention that LRU II, where the remote cache is allowed to update its local storage upon a global cache hit, achieves higher LHR than LRU I across the considered range of β. This is due to the fact that such remote cache updates provide an indirect and limited, yet useful indication

of global statistics, whose exploitation leads to higher LHR for the remote cache. On the other hand, we don't notice major performance differences between the variants of the statistics-based caching schemes.

- GHR: *LRU outperforms significantly the other caching schemes, especially for moderate and large values of* β, with the exception of SG-C II which performs much better than LRU for $\beta = 0.99$. The superiority of LRU against the statistics-based strategies is explained by the fact that the remote cache acts as a second-level cache that deals with requests that have been filtered by the target cache and, therefore, do not follow a Zipf distribution. Similarly, the random selection of cached objects whose score controls whether an insertion of an external object into the cache will take place or not, makes SG-C to perform similar to LRU—or even better than LRU when the remote cache is allowed to update its scores. Furthermore, we see that the GHR of LRU improves as β gets larger, in contrast to the behavior of the remaining caching strategies. This phenomenon is attributed to the filtering of the requests from the target cache, which reduces the influence of content popularity in caching and replacement decisions. Another interesting observation is that the corresponding variants of the statistics-based caching schemes perform close to each other, as well as that *in the majority of cases the exploitation of global statistics improves the GHR.*
- THR: *The THR of all caching schemes improves as* β *increases.* Furthermore, we see that *the statistics-based caching techniques outperform only slightly LRU*, with their performance gain over LRU being a little bit higher for larger values of β. This is because their LHR gains are partially compensated by the GHR gains of LRU. Also, we observe that these statistics-based strategies perform close to each other, with the notable exception of SG-C I whose THR drops significantly when $\beta = 0.99$ (in comparison to other similar strategies).

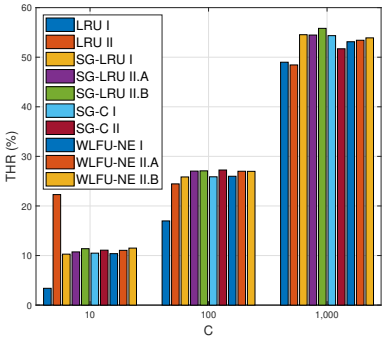
Use Case (2): We set $\beta = 0.75$ and vary the cache size as $C = \{10; 100; 1,000\}$ (i.e., as $\{0.1\%; 1\%; 10\%\}$ of the catalog size). We should mention that a cache size equal to the 10% of the catalog size is rather unrealistic and it has been added here only as a means to study the cache behavior in extreme conditions. The hit rates are shown in Fig. 4.

- LHR: *The LHR of all caching strategies grows with the cache size*, as expected. The statistics-based caching schemes perform close to each other, with the exception of the case for $C = 1,000$ where the performance of the WLFU-NE variants and of SG-C II is a little worse and considerably worse, respectively, than the performance of the SG-LRU variants and SG-C I. The small performance degradation of WLFU-NE for large cache size is caused mainly by the fact that the successive neighbor exchange updates take a long time to produce a sorted cache list w.r.t. the score of the cached objects. On the other hand, the significant performance degradation of SG-C II for large caches is attributed to the fact that such caches store along



(a) Local hit rates.

(b) Global hit rates.

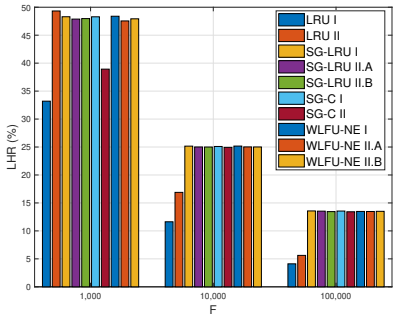


(c) Total hit rates.

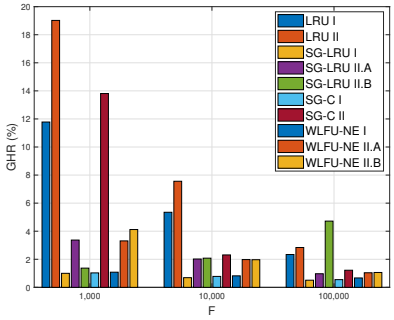
Fig. 4: Cache hit rates for varying cache size C.

with the few very popular objects a large number of not so popular objects. In this case, the comparison of an external object's score with the score of a pretty-much randomly selected cached object is highly inefficient. Also, for the same reason the update of the remote cache's scores upon a global cache hit further decreases its LHR. *The statistics-based caching methods outperform LRU, except for small caches where LRU II (which constantly outperforms LRU I) performs better*. This is because the influence of request count statistics on the caching efficiency is smaller for small caches.

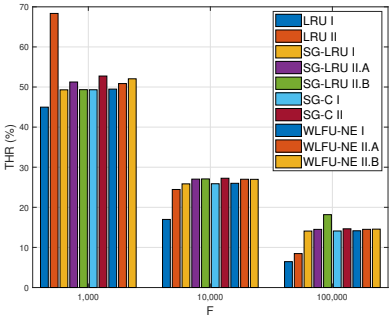
– GHR: *LRU outperforms significantly the statistics-based caching schemes, especially for small caches*, with the exception of SG-C II for $C = 1,000$. Also, *variants II perform in general better than variants I*, with few exceptions when $C = 1,000$.
– THR: *The THR of all caching schemes improves as C increases (less aggressively for LRU when C is small). LRU II outperforms the statistics-based caching strategies only for $C = 10$*. The statistics-based caching vari-



(a) Local hit rates.



(b) Global hit rates.



(c) Total hit rates.

Fig. 5: Cache hit rates for varying catalog size F.

ants perform similar to each other. In general, *variants II present a little higher THR than variants I*, with the exception of SG-C for the case where $C = 1,000$ due to the fact that its high GHR is compensated by its low LHR. SG-LRU II.B performs slightly better than the other statistics-based caching schemes across the whole range of C.

Use Case (3): We fix $C = 100$ and vary the catalog size, so that the cache size C is the $\{10\%; 1\%; 0.1\%\}$ of the catalog size, as in Use Case (2), i.e., $F = \{1,000; 10,000; 100,000\}$. The hit rates are shown in Fig. 5. We observe similar behavior with the one illustrated in Fig. 4 for the corresponding catalog size / cache size ratios, with a few exceptions that we will present here.

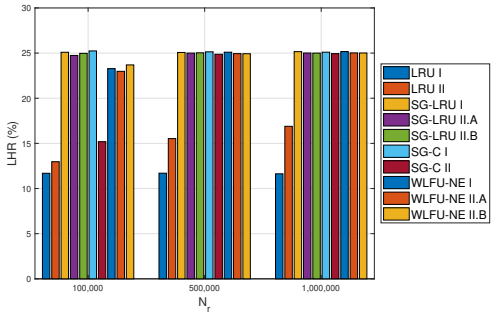
- LHR: The performance of WLFU-NE does not present any degradations, since the cache size is moderate. Also, LRU II presents a small performance gain over the statistics-based caching strategies for small catalog size.
- GHR: SG-LRU II.B performs better than LRU for large catalog size / cache size ratio. Moreover, SG-LRU II.A presents a small performance gain over

SG-LRU II.B for small catalog size, whereas for large catalog size we see the opposite phenomenon.
– THR: LRU II outperforms significantly the statistics-based caching schemes for small catalog size. SG-C II and SG-LRU II.B achieve the highest THR among the statistics-based caching methods for small and large catalog size, respectively.

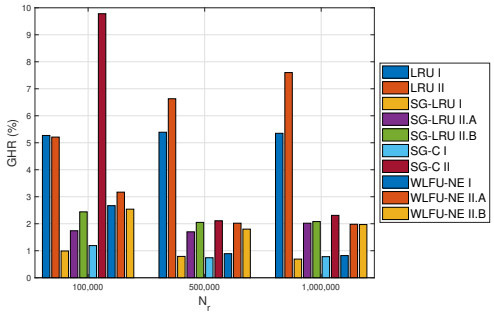
Use Case (4): We set $F = 10,000$ and vary the number of user requests as $N_r = \{100,000; 500,000; 1,000,000\}$. The hit rates are depicted in Fig. 6. We note that SG-LRU and LRU I have almost reached their steady-state LHR already after $100,000$ requests, while SG-C and WLFU-NE needed $500,000$ requests due to the random selection of the least valuable cached object and the successive cache updates based on NE, respectively. Also, we see that LRU II converged after $1,000,000$ requests due to the filtering of the requests seen by the remote cache. The same picture holds true for the GHR. Regarding the THR, we see that LRU I and the statistics-based caching strategies have approached their steady-state performance after $100,000$ requests, whereas LRU II needed $1,000,000$ requests.

Use Case (5): We set the number of requests to $N_r = 1,000,000$ and vary the window size of the statistics-based caching schemes from $W = 1$ up to $W = 1,000,000$, i.e., from the 0.0001% up to the 100% of the number of user requests. The hit rates are depicted in Fig. 7. We note that the LHR and THR of all caching strategies increases as W grows while the GHR decreases. This is because (i) in larger windows the request frequency information influences more the caching and replacement decisions than the request recency information, thus improving the LHR and decreasing the GHR; and (ii) the LHR gains compensate for the GHR losses. For small window size, the statistics-based caching strategies present LRU-like behavior and they start to approach their baseline performance that is achieved for $W = 100,000$ for moderate window sizes, with WLFU-NE having a slower convergence rate than the other caching methods. SG-LRU II.B constitutes an exception, since its performance is relatively high even for $W = 1$. When the window size becomes practically unlimited (i.e., equal to the number of requests), the hit rates remain constant or vary slightly, except for SG-C I whose LHR and GHR drops and increases significantly, respectively.

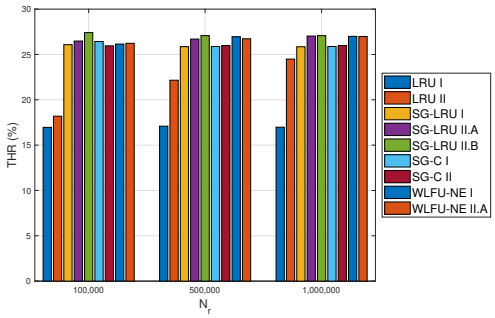
Summary: The statistics-based strategies achieve higher LHR and THR than LRU (except for small caches), with the best performance obtained for large values of β, C, and W. SG-C constitutes an exception, since for large values of β or C or for practically infinite window size, it presents LRU-like behavior (like most statistics-based methods do for small window size). LRU, on the other hand, outperforms these caching schemes in terms of GHR. In most cases, the use of global statistics improves the GHR. We should note that the caching methods that achieve high LHR are preferable over caching schemes



(a) Local hit rates.



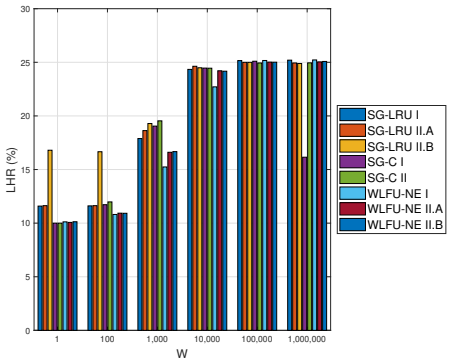
(b) Global hit rates.



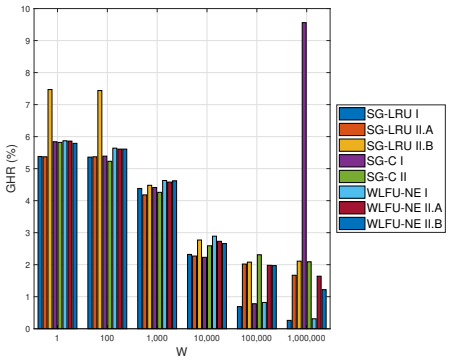
(c) Total hit rates.

Fig. 6: Cache hit rates for varying number of user requests N_r.

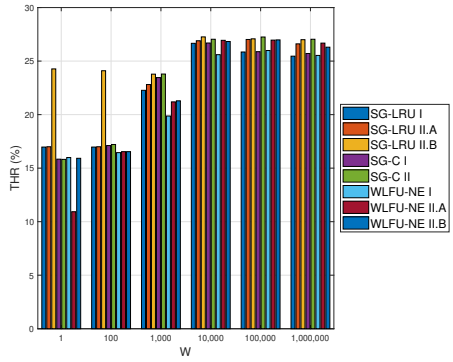
that achieve high GHR and have comparable THR with them, since local caching results in higher delay reduction and network traffic savings.



(a) Local hit rates.

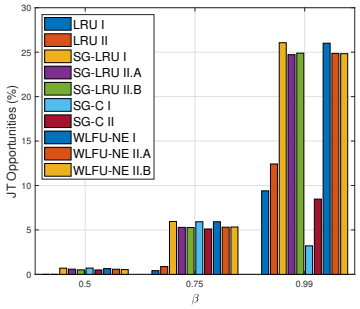


(b) Global hit rates.

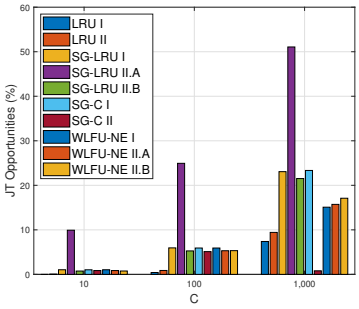


(c) Total hit rates.

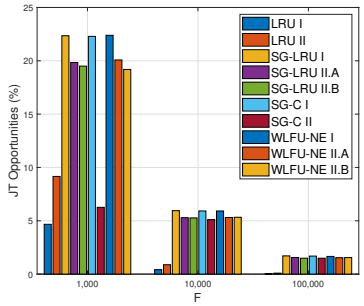
Fig. 7: Cache hit rates for varying window size W.



(a) Varying Zipf shape parameter β.



(b) Varying cache size C.



(c) Varying catalog size F.

Fig. 8: Percentage of cache-aided JT opportunities for varying β, C, and F.

6.2 Joint Transmission Opportunities

Next, we study the ability of the considered caching strategies to create JT opportunities in the aforementioned use cases. Note that if MS_{11} (the sole active user associated with BS_1) requests O_i and MS_{12} (the sole active user associated with BS_2) requests O_j $(i,\ j \in \mathcal{F})$, then each one of C_1 and C_2 should have stored both these objects, for cache-aided JT to take place.

In Fig. 8(a) we see that the percentage of JT opportunities grows with β for all caching schemes, as expected. This is because the caching efficiency improves for higher values of β and the considered caching methods exploit (in a bigger or smaller extent) content popularity information in the caching and replacement decisions, thus ending up often with similar cache lists. SG-C I constitutes an exception, since the performance of this caching strategy drops for $\beta = 0.99$. This is due to its LHR loss in this scenario that is caused by the random selection of the candidate for eviction cached object. Naturally, the number of JT opportunities is small for low values of β and all caching strategies perform close to each other in this scenario. For moderate and high values of β, the statistics-based caching methods outperform significantly LRU

(with the exception of SG-C), thanks to the direct exploitation of request frequency information.

In Fig. 8(b) we note that the number of JT opportunities increases with the cache size, as expected. This is because the caching efficiency improves for larger cache sizes and the probability of finding objects that have been stored in both caches increases for larger cache lists. Again, SG-C II represents an exception, since its performance drops for $C = 1,000$. The statistics-based caching strategies outperform significantly LRU, with the higher relative and absolute performance gain noticed for small and large cache sizes, respectively. Moreover, we observe that SG-LRU II.A presents the best performance across the whole relevant range of C. Similar observations are obtained from Fig. 8(c), where the catalog size is varied, for equal catalog size / cache size ratios.

As depicted in Fig. 9(a), LRU II, SG-LRU, and SG-C I approach their baseline performance that is achieved for $N_r = 1,000,000$ already when $N_r = 100,000$, while WLFU-NE and SG-C II require $500,000$ requests (in fact, for $100,000$ requests the performance of SG-C II is LRU-like) and LRU I needs $1,000,000$ requests.

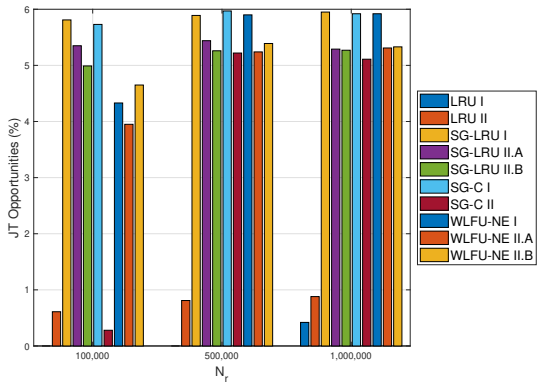
Finally, in Fig. 9(b) is shown that the performance of all statistics-based strategies improves with the window size up to $W = 100,000$ and then remains constant or increases slightly for $W = 1,000,000$, with the exception of SG-C I whose performance degrades significantly. We note that for small window size, all caching strategies present an LRU-like behavior. From $W = 1,000$ and onwards, their performance starts to improve, with WLFU-NE presenting a slower rate of improvement in comparison to the other caching schemes. The best performance is achieved for large window size by SG-LRU I, SG-C I (with the exception of the scenario where $W = 1,000,000$), and WLFU-NE I.

In summary, we note that the statistics-based strategies outperform LRU and can create a significant number of cache-aided JT opportunities for moderate and large values of β, C, and W, especially for caches that deal with a large number of user requests.

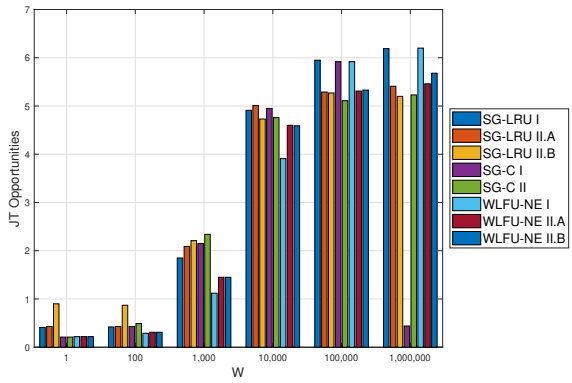
6.3 Sum Spectral Efficiency

Then, we study the performance of the considered resource allocation techniques for the CoMP-enabled underlay spectrum sharing setup depicted in Fig. 1, assuming $M = 2$, $K = 2$, and $N = 4$ and starting from the CoMP-CBF case. We are interested in plotting the average sum-SE that is achieved after $N_{\text{sim}} = 1,000$ simulation runs $\bar{R} = \mathbb{E}(R)$ (where the expectation is taken over the channel realizations) vs. the average SNR $\bar{\gamma}$. We consider the application of ZF precoding, unless it is explicitly stated otherwise.

First, we vary the IPT as $P_I = \{30; 20\}$ dB and the transmission power of the PS as $P = \{0; 10\}$ dB and compare the performance of C-ICPA and C-IUPA. In Fig. 10(a) we see that the average sum-SE increases with $\bar{\gamma}$, as expected. C-ICPA performs close to C-IUPA for relaxed P_I and small P but



(a) Varying number of requests N_r.



(b) Varying window size W.

Fig. 9: Percentage of cache-aided JT opportunities for varying N_r and W.

its average sum-SE is reduced significantly as P gets larger and starts to floor faster as the IPT becomes harder.

Next, we consider a scenario where $P_I = 30$ dB and $P = 0$ dB and study the performance of CQA-ICPA for a group of users QoS_1 with QoS (minimum rate) requirements $\left[0.10\ 0.10\ 0.10\ 0.10\right]\bar{\gamma}$ and for another group of users QoS_{uni} where the QoS requirement of each user follows the uniform distribution $\mathcal{U}(0.10\bar{\gamma}, 0.50\bar{\gamma})$. We note that CQA-ICPA performs very close to C-ICPA for the less demanding group of users QoS_1 (the sum-SE starts to drop only at high SNR) and just above C-ICEPA for the more demanding group QoS_{uni}.

In Fig. 10(c) is depicted the performance of C-IUPA and C-ICPA under the scenario where $P_I = 30$ dB and $P = 0$ dB for varying number of antennas $N = \{4; 8\}$. We note that the increase in the number of transmit antennas (or,

better, the increase of the ratio of BS antennas over MSs) results in higher average sum-SE, but the performance trends are the same in both cases.

Fig. 10(d) illustrates the performance of C-ICPA when RZF, ZF, and MRT is employed vs. the performance of its C-IUPA counterparts for a scenario where $P_I = 30$ dB, $P = 0$ dB, and $N = 4$. We notice that all C-ICPA schemes perform close to the corresponding C-IUPA methods. Furthermore, we note that RZF precoding outperforms ZF precoding, presenting higher average sum-SE gains for low and moderate average SNR, as well as that the sum-SE of MRT floors, regardless of whether C-ICPA or C-IUPA is applied, due to the existence of uncoordinated intra-system CCI within the cellular network, as expected.

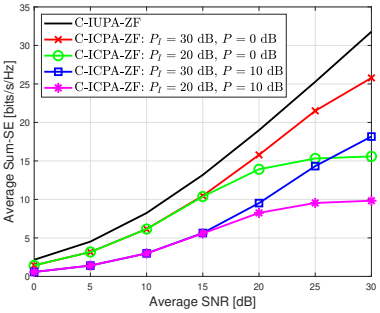
In Fig. 10(e) and Fig. 10(f) is depicted the performance of C-ICPA, CQA-ICPA for QoS_1 and for QoS_{uni}, and C-ICEPA when coordinated ZF precoding is applied vs. the performance of their C-IUPA counterparts when coordinated P-ZF precoding is utilized for the scenarios where $\{P_I, P\} = \{5 \text{ dB}, 0 \text{ dB}\}$ and $\{P_I, P\} = \{0 \text{ dB}, 0 \text{ dB}\}$, respectively. We note that while the performance of the C-ICPA variants floors quickly for hard IPT, the combination of C-IUPA variants and P-ZF precoding provides a substantial improvement of the average sum-SE in such scenarios and is unaffected by the IPT value, as expected.

In summary: C-ICPA with ZF precoding approaches its C-IUPA counterpart for relaxed IPTs and small RIS CCI. CQA-ICPA achieves a performance in between of that of C-ICPA and ICEPA. The performance of these schemes improves for more BS antennas or when RZF precoding is applied instead of ZF precoding. Finally, the use of C-IUPA variants with P-ZF precoding improves substantially the performance for hard IPTs.

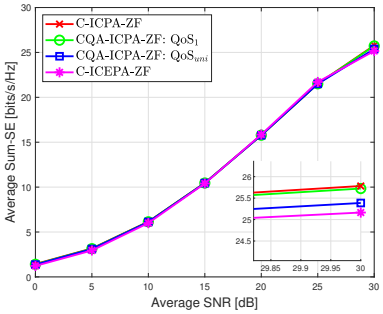
In the last test, we consider a setup where $M = 2$, $K = 1$, and $N = 4$, and, assuming the use of SG-LRU I under a scenario where $\beta = 0.99$, $C = 100$, $F = 10,000$, $N_r = 1,000,000$, and $W = 100,000$, we compare the performance of CoMP-CBF vs. the performance of a hybrid approach where CoMP-CBF is utilized only when cache-aided CoMP-JT cannot take place (that is, about 74% of the time for this caching scenario) after $N_{\text{sim}} = 1,000$. The application of coordinated ZF precoding and C-ICEPA is considered in both cases. We note in Fig. 11 that the hybrid method slightly outperforms CoMP-CBF for relaxed IPT and performs slightly worse for more tight IPT values. The main outcome of this test, though, is that it is possible to use cache-aided CoMP-JT in spectrum sharing setups, in order to further improve the QoS of the cell-edge users.

7 Summary and Conclusions

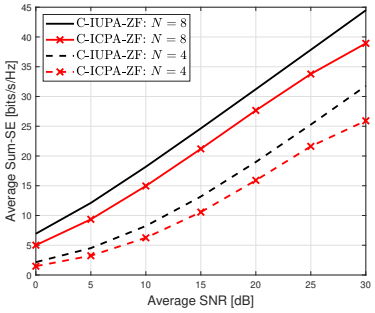
In this work, we presented a coordinated caching strategy and a number of statistics-based caching schemes. The latter achieve higher LHR and THR than the "de facto" LRU caching method and create more JT opportunities, especially for large β, C, and W, whereas LRU achieves better GHR. The



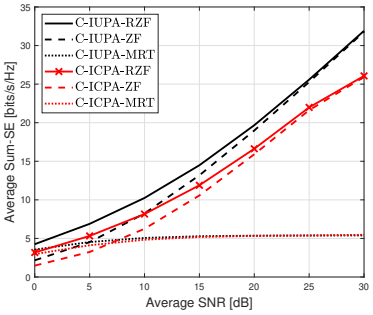
(a) C-IUPA vs. C-ICPA for varying $\{P_I, P\}$.



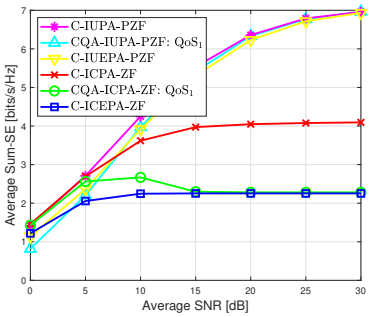
(b) C-ICPA vs. CQA-ICPA vs. ICEPA.



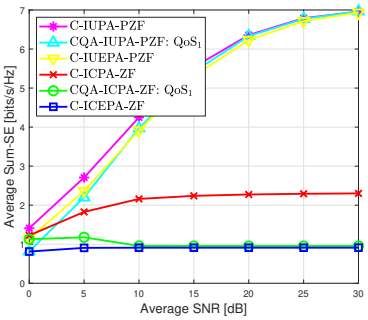
(c) C-IUPA vs. C-ICPA for varying N.



(d) RZF vs. ZF vs. MRT.



(e) ZF vs. P-ZF for $\{P_I, P\} = \{5 \text{ dB}, \ 0 \text{ dB}\}$.



(f) ZF vs. P-ZF for $\{P_I, P\} = \{0 \text{ dB}, \ 0 \text{ dB}\}$.

Fig. 10: Average sum-SE vs. average SNR.

use of cooperative caching variants that utilize global statistics improves the GHR.

Moreover, we proposed coordinated resource allocation techniques based on linear precoding schemes for maximizing the sum-SE of CoMP-enabled networks in underlay spectrum sharing setups in either a QoS-aware or QoS-

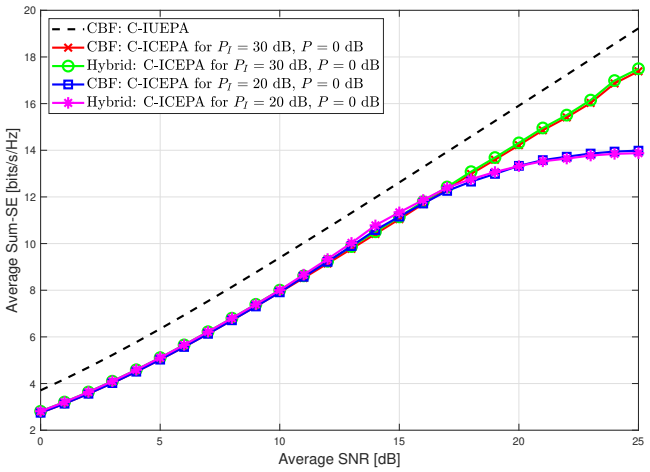


Fig. 11: CoMP-CBF vs. hybrid CoMP-CBF / cache-aided CoMP-JT.

agnostic context. The simulation results revealed that the combination of these methods performs close to the interference-unconstrained techniques for any IPT and that the performance is improved for more relaxed IPT, more transmit antennas, or by using RZF precoding instead of ZF. Finally, we presented a hybrid CoMP-CBF / cache-aided CoMP-JT approach.

References

1. E. Bjornson, J. Hoydis, and L. Sanguinetti, "Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, Nov. 2017.
2. IEEE SIG Cognitive Radio in 5G, "Novel Spectrum Usage Paradigms for 5G," White Paper, Nov. 2014.
3. 5GPPP, "5G Vision," White Paper, 2015.
4. E. Bjornson and E. Jorswieck, "Optimal Resource Allocation in Coordinated Multi-Cell Systems," *Foundations and Trends in Communications and Information Theory*, vol. 9, no. 2-3, pp. 113–381, Jan. 2013.
5. E. Castaneda *et al.*, "An Overview on Resource Allocation Techniques for Multi-User MIMO Systems," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 239–284, 2017.
6. I. F. Akyildiz *et al.*, "A Survey on Spectrum Management in Cognitive Radio Networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 40–48, Apr. 2008.
7. S. Pandit and G. Singh, "An Overview of Spectrum Sharing Techniques in Cognitive Radio Communication System," *Wireless Networks*, vol. 23, no. 2, pp. 497–518, Feb. 2017.
8. F. Mehmeti and T. Spyropoulos, "Performance Analysis, Comparison, and Optimization of Interweave and Underlay Spectrum Access in Cognitive Radio Networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7143–7157, Aug. 2018.
9. WG FM53, "ECC Report 205: Licensed Shared Access," Tech. Rep., Feb. 2014.

10. A. Morgado *et al.*, "Dynamic LSA for 5G Networks: The ADEL Perspective," in *European Conference on Networks and Communication (EuCNC)*, Paris, France, 2015, 29 June - 2 July.
11. ETSI Technical Committee Reconfigurable Radio Systems (TC RRS), "ETSI TR 103 588 V1.1.1: Feasibility Study on Temporary Spectrum Access for Local High-Quality Wireless Networks," Tech. Rep., Feb. 2018.
12. L. Gallo *et al.*, "Weighted Sum Rate Maximization in the Underlay Cognitive MISO Interference Channel," in *22nd International IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Toronto, ON, Canada, 2011, pp. 661–665, 11-14 September.
13. D. Denkovski *et al.*, "Generic Multiuser Coordinated Beamforming for Underlay Spectrum Sharing," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2285–2298, Jun. 2016.
14. G. Hasslinger and K. Ntougias, "Evaluation of Caching Strategies Based on Access Statistics of Past Requests," in *17th International GI/ITG MMB and DFT Conference*, ser. Lecture Notes in Computer Science (LNCS), Springer, Ed., vol. 8376, Bamberg, Germany, Mar. 2014, pp. 120–135.
15. G. Hasslinger, K. Ntougias, and F. Hasslinger, "A New Class of Web Caching Strategies for Content Delivery," in *16th International Telecommunications Network Strategy and Planning Symposium (Networks)*, Funchal, Portugal, 2014, 17-19 Sept.
16. K. Shanmugam *et al.*, "Femtocaching: Wireless Video Content Delivery Through Distributed Caching Helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
17. F. Zhou *et al.*, "A Cache-Aided Communication Scheme for Downlink Coordinated Multipoint Transmission," *IEEE Access*, vol. 6, pp. 1416–1427, Dec. 2017.
18. A. Tuholukova *et al.*, "Optimal Cache Allocation for Femto Helpers with Joint Transmission Capabilities," in *IEEE International Conference on Communications (ICC)*, Paris, France, May 2017.
19. W. C. Ao and K. Psounis, "Fast Content Delivery via Distributed Caching and Small Cell Cooperation," *IEEE Transactions on Mobile Computing*, vol. 17, no. 5, pp. 1048–1061, May 2017.
20. G. Hasslinger *et al.*, "Performance Evaluation for New Web Caching Strategies Combining LRU with Score Based Object Selection," *Computer Networks*, vol. 125, pp. 172–186, Oct. 2017.
21. M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," in *ACM SIGCOMM Conf. Internet Measurement (IMC)*, San Diego, CA, USA, 2007, pp. 1–14, 24-26 October.
22. T. Qiu *et al.*, "Modeling Channel Popularity Dynamics in a Large IPTV System," in *SIGMETRICS International Joint Conference on Measurement and Modeling of Computer Systems*, Seattle, WA, USA, Jun. 2009, pp. 275–286.
23. K. Ntougias *et al.*, "Simple Cooperative Transmission Schemes for Underlay Spectrum Sharing Using Symbol-Level Precoding and Load-Controlled Arrays," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019, to appear.
24. S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
25. L. Breslau *et al.*, "Web Caching and Zipf-like Distributions: Evidence and Implications," in *18th IEEE Infocom*, New York, NY, USA, 21-25 March 1999, pp. 126–134.
26. R. Bolla *et al.*, "A Measurement Study Supporting P2P File-Sharing Community Models," *Computer Networks: Special Issue on Content Distribution Infrastructures for Community Networks*, vol. 53, no. 4, pp. 485–500, March 2009.
27. C. Fricker, P. Robert, and J. Roberts, "A Versatile and Accurate Approximation for LRU Cache Performance," in *24th International Teletraffic Congress (ITC)*, Krakow, Poland, 2012.
28. G. Hasslinger *et al.*, "Comparing Web Cache Implementations for Fast O(1) Updates Based on LRU, LFU and Score Gated Strategies," in *23rd IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, Barcelona, Spain, 2018, 17-19 Sept.