Distributed Caching and Small Cell Cooperation for Fast Content Delivery

Weng Chon Ao
University of Southern California
wao@usc.edu

Konstantinos Psounis
University of Southern California
kpsounis@usc.edu

ABSTRACT

Modern wireless devices such as tablets and smartphones are pushing the demand for higher and higher wireless data rates. The vast majority of this demand comes from media content. In this paper we propose to combine two recent ideas, distributed caching of content in small cells, and, cooperative transmissions from nearby base stations/BSs (generally known as coordinated multi-point), to achieve unprecedented content delivery speeds while reducing backhaul cost and delay. A key characteristic of our architecture is the interdependence between the caching strategy and the PHY/MAC layer coordination. Specifically, the caching strategy may cache different content in nearby BSs to maximize the hit ratio, or cache the same content in multiple nearby BSs such that the corresponding BSs can transmit concurrently, e.g. to multiple users using zero force beamforming, and achieve multiplexing gains. With this in mind, given the popularity distribution of the content, the available cache size, and the network topology, we devise optimal strategies of caching such that the throughput of the system is maximized. Our analytical and simulation results show that our system yields significantly faster content delivery, which, under realistic scenarios and assumptions can be one order of magnitude faster than that of legacy systems.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]: Wireless Communication

General Terms

Performance, Theory

Keywords

Heterogeneous cellular networks, Caching, Coordinated Multi-Point, Multi-user MIMO, Diversity, Multiplexing

1. INTRODUCTION

This work was supported by NSF grant ECCS-1444060.

MobiHoc'15, June 22–25, 2015, Hangzhou, China.

http://dx.doi.org/10.1145/2746285.2746300.

The proliferation of advanced mobile devices such as smart phones and tablets together with the popularity of video streaming are causing a tremendous growth of data traffic in cellular networks. To address this challenge the cellular industry is advocating the deployment of small cells (via the use of low power nodes like micro-BSs, pico-BSs and femto-BSs) together with traditional macro-BSs in a heterogeneous networking architecture [6, 10]. These low power nodes provide short-range localized communication links resulting in a higher density of spatial reuse of radio recourses and thus in higher overall network throughput.

Deploying a dense network of low power nodes comes with its own challenges. One such challenge that service providers consistently rank high is the deployment cost associated with connecting all the small cells to the backbone with fast links. Motivated by this, there is a growing interest to cache popular content to those low power nodes in a distributed manner, effectively trading off fast backhaul capacity with storage capacity. Specifically, the authors in [12, 13, 24] have introduced the concept of FemtoCaching, which is the idea of embedding femto-BSs with high storage capacity to store popular video files. When a user requests a video file, the user may be served by a nearby femto-BS that has the requested file in its cache over a high rate short-range wireless link. If the requested file is not in the cache of any nearby femto-BS, the user will be served directly by the macro-BS over a low rate long-range wireless link. Since the popularity distribution of video files changes at a much slower pace than that of user requests, cache updates (downloading popular video files via backhaul into the caches) can be done at off-peak hours, which results in a significant reduction of backhaul cost and delay while maintaining the performance benefits of a dense deployment of low power BSs.

A dense deployment of low power BSs yields even higher throughput when multiple neighboring BSs coordinate their data transmissions such that they aggregate constructively, see, for example, two recent tutorial-style papers [16, 18] and [2,3,23] where real-world testbeds have been created and tested using software defined radios. As a matter of fact, in the absence of such BS coordination, interference between nearby BSs may cancel the performance gains of dense deployments, and service providers consistently rank the technological challenges related to this issue as yet another major challenge in the deployment of small cells. There are many schemes for BS coordination, and in this paper we will consider the two most basic/popular ones: Maximum Ratio Transmission (MRT) and Zero-Forcing BeamForming

(ZFBF) [25]. Consider that low power BSs form cooperation clusters. Then, under MRT, each BS in the cooperation cluster beamforms to a user such that the signals from the neighboring BSs are coherently combined resulting in a diversity gain [4]. Under ZFBF, the BSs in the cluster are simultaneously transmitting multiple data streams to multiple users [3,9], resulting in a multiplexing gain. Note that in the absence of offline cache updates, both MRT and ZFBF would further increase the cost and delay associated with backhaul, as they require multiple copies of the same files to be distributed to multiple BSs.

In this paper, we propose to combine FemtoCaching and femto-BS cooperation. The proposed cooperation scheme is cache/data-driven in the sense that if a typical user makes a request of a video file, only the neighboring femto-BSs that have the requested video file in their caches will participate in the cooperative transmission. In other words, the cluster of cooperating femto-BSs is dynamically formed on a per-request basis. An important aspect of our framework is the joint optimization of the cache allocation (content placement) in the application layer and the cooperative transmission techniques (MRT for diversity and ZFBF for multiplexing) in the physical layer. We jointly optimize these aspects of the system because caching different content in nearby caches increases hit ratio, but caching the same content increases the chances to get diversity and multiplexing gains. In general, the optimal cache allocation depends on a number of parameters, including the file popularity distribution, the cache size, the number of neighboring femto-BSs, and the transmission rate of the macro-BS in comparison to that of a femto-BS.

The remainder of this paper is organized as follows. We present related work in Section 2. Section 3 describes the setup, the caching strategies, and the cache-driven cooperation policies. In Section 4 we derive analytical formulas for the achieved rates under a variety of scenarios considering both background noise and co-channel interference. In Section 5 we present numerical results for a number of real-world scenarios, highlighting the rate gains from our framework. Notably, our schemes can increase the speed of content delivery by an order of magnitude without requiring fast backhaul speeds. Last, Section 6 discusses practical considerations and Section 7 concludes the paper.

2. PRIOR WORK AND CONTRIBUTIONS

This work touches upon a number of prior lines of work. The setup is that of heterogeneous networks, formed by distributing multi-tier low power nodes (e.g., micro-BSs and femto-BSs) in macro-cellular networks, see, for example, two recent tutorial-style papers and references therein [6, 10]. It is building upon prior work on BS cooperation, more generally known as Coordinated Multi-Point (CoMP), and FemtoCaching. There is a long line of research in CoMP, see, for example, [16, 18]. FemtoCaching has been recently introduced to trade off backhaul capacity with cache capacity [12, 13, 24] and can be further applied to device-to-device communication networks [11,17] and to coded caching [21, 22]. FemtoCaching itself is building up upon prior work on distributed caching, content placement schemes and content distribution networks, see, for example, [5].

In addition to using standard analytical tools like convex and integer optimization, combinatorics, and Shannon rate formulas, we also use stochastic geometry [14] to take into account co-channel interference in the context of heterogeneous networks, see, for example, the relevant analysis in [7].

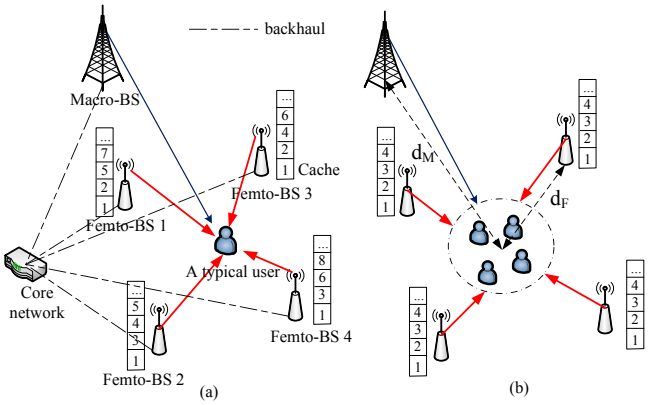


Figure 1: System model for cache-driven femto-BS cooperation: (a) randomized caching and MRT (b) threshold-based caching and ZFBF.

Directly related to this work is [19, 20] where the authors use a coding scheme to introduce redundancy in caches and create CoMP opportunities for cooperative transmissions. A fundamental difference between this prior work and our paper is that we consider the effect of cache misses, since any type of redundancy decreases the number of distinct files that can be stored in finite size caches. To optimize the system performance, we appropriately control the stored redundancy for each individual file and dynamically (per-request basis) form a cluster of cooperating femto-BSs.

Our contributions are as follows: We combine the concepts of FemtoCaching and BS cooperation to propose a novel, high-performing system architecture. We derive analytical expressions for the user rates and jointly optimize the caching strategy and the PHY layer cooperation. We devise efficient caching strategies for providing diversity gains under MRT, multiplexing gains under ZFBF, and the optimal diversity-multiplexing tradeoff. Last, we study the performance of our schemes under practical scenarios and address deployment considerations.

3. SYSTEM MODEL

3.1 Topology

Consider a typical user in a macrocell. Suppose there are N femto-BSs and another $K-1$ users in the neighborhood of the typical user. We denote by $d_{0,k}$, $1 \leq k \leq K$, and $d_{j,k}$, $1 \leq j \leq N$, $1 \leq k \leq K$, the distance between the macro-BS and the kth user, and the jth femto-BS and the kth user respectively. Let these K co-located users be associated with the same N neighboring femto-BSs and the macro-BS. These N neighboring femto-BSs are candidates for cooperative transmissions, see Fig. 1a for a scenario where femto-BSs transmit the same content (say, file 1) to one user, and Fig. 1b for a scenario where femto-BSs transmit concurrently to multiple users (say, to four users four different files, namely file 1, 2, 3, and 4).

In a typical real-world scenario one may have tens or hundreds of femto-BSs inside a macrocell and hundreds or thousands of users. Thus, femto-BSs would be grouped into clusters of nearby femto-BSs which can concurrently serve a number of users. For example, one may have one such clus-

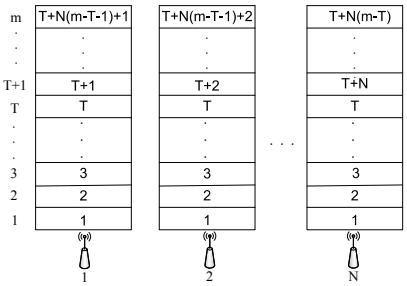


Figure 2: Caching strategy under threshold-based caching and ZFBF.

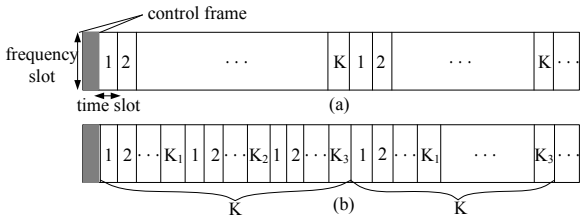


Figure 3: Control and data frames under (a) randomized caching and MRT, (b) threshold-based caching and ZFBF.

ter per floor on a large building or one cluster per building. In the rest of this paper we will focus on one such cluster of co-located femto-BSs and users.[1]

In general, the users, the macro-BS, and the femto-BSs would be assigned different time-frequency slots (resource blocks) for data transmissions. We will only focus on the downlink, and thus on transmissions towards the K users. Also, for simplicity and without loss of generality, we will focus on a single frequency slot which the macro-BS and the femto-BSs may use to transmit to the users. In some of the physical layer schemes that we study, one time slot can be used for a single user only, e.g. MRT, while in others it can be used for multiple users concurrently, e.g. ZFBF. Depending on whether a user is served by the macro-BS, a single femto-BS, or multiple femto-BSs using MRT or ZFBF, the data rate that it will receive during this time slot will vary. Last, a control frame will be used to collect requests from the users at the macro-BS, and then the users will be served in subsequent time slots (see Fig. 3). Since a request for a media file corresponds to multiple data packets, such control frames will be infrequent.

3.2 Caching Strategies and Cache-driven Cooperation policies

Suppose that there is a library of M video files which are ordered according to their (normalized) popularity p_i, $1 \leq i \leq M$, $p_1 \geq p_2 \geq \cdots \geq p_M$, $\sum_{i=1}^{M} p_i = 1$. In other words, a typical user would request the ith file with probability p_i. The macro-BS stores (or has access to) all of the M files, while each femto-BS has a cache that can store on average up to m files with $m < M$. We have the following two different caching and cooperation strategies aiming at providing diversity gain and multiplexing gain.

3.2.1 Providing a diversity gain by randomized caching and MRT

For each femto-BS the caching strategy caches the ith file with probability q_i, $0 \leq q_i \leq 1$, $i = 1, \cdots, M$, subject to the probabilistic cache size constraint $\sum_{i=1}^{M} q_i \leq m$. (To simplify the analytical exposition and without loss of generality, here we have assumed that all files have the same size, the unit of storage/size is a file, and the average cache size is m.) Note that q_i, $i = 1, \cdots, M$ are design parameters.

Then, the cooperation policy dictates that for a request of the ith file, generated by a typical user according to the popularity distribution law p_i, the request will be jointly served by the femto-BSs that have the requested file in their caches. Thus, no data file exchanges are needed for the cooperation.[2] Under MRT, each femto-BS that has the requested file in its cache beamforms its signal to the typical user so that the signals (from the cooperating femto-BSs) are coherently combined at the receiver, producing a diversity gain for the desired signal. If none of the femto-BSs have the requested file, the request will be served by the macro-BS. For example, in Fig. 1a, if a user requests the 2nd file, then femto-BSs 1 and 3 will beamform to the user. Fig. 3a shows a typical sequence of time slots under MRT, where each of the K users receives its data at one of K time slots.

3.2.2 Providing a multiplexing gain by threshold-based caching and ZFBF

Aiming at providing multiplexing gain we consider the following caching strategy. We first choose a threshold T, where $0 \leq T \leq m$, which is a design parameter (we will generalize to the case with multiple thresholds in Section 4.3). We cache the files 1 to T in all of the femto-BSs, and cache the files $T+1$ to $T+N(m-T)$ in exactly one of the femto-BSs. That is, we have N copies for each of the most popular files 1 to T, one copy of the files $T+1$ to $T+N(m-T)$, and the remaining files $T+N(m-T)+1$ to M can be downloaded only by the macro-BS (see Fig. 2).

We associate the K users with K different time slots as in Fig. 3b. The K users can be partitioned into three groups according to their requests: The first group consists of users with requests of files between 1 and T, i.e., for files that are cached in all femto-BSs (the group size is denoted as K_1), the second group consists of users with requests of files between $T+1$ and $T+N(m-T)$, i.e., for files that are cached in only one femto-BS (the group size is K_2), and the third group consists of users with requests of files between $T+N(m-T)+1$ and M, i.e., for files that can be fetched only by the macro-BS (the group size is $K_3 = K - K_1 - K_2$). Since all the femto-BSs cache the files requested by users of the first group, the N femto-BSs can coordinate their transmissions when serving these K_1 users. Specifically, the N femto-BSs can simultaneously serve these K_1 users ($K_1 \leq N$) during K_1 time slots using ZFBF producing a multiplexing gain of order K_1.[3] No cooperative transmission is used for serving a user of the second (third) group because only a single femto-BS (the macro-BS) has the user's requested file. Note that if one desires all users to receive similar rates, one may assign less than K_1 slots to the users of the first group, e.g.

[1] It is beyond the scope of this paper to further investigate clustering algorithms.

[2] Note that control signals such as channel state information may be required to be distributed among the femto-BSs in the cooperation cluster.

[3] By using ZFBF, suppose there are N transmitters and K receivers ($K \leq N$), the N transmitters can transmit K independent (non-interfering or spatially isolated) streams to the K receivers simultaneously, each with a diversity of order 1.

Table 1: Main notation

Number of femto-BSs	N
Distance between a typical user and a macro(femto)BS	d_M (d_F)
Number of video files	M
Cache size in a femto-BS	m
Number of users	K
File popularity distribution	p_i (cdf v_i)
Probability of caching the ith file for MRT	q_i
Caching threshold for ZFBF	T
Bandwidth of a frequency slot	W
Transmit power of macro(femto)-BS	P_M (P_F)
Data rate of macro(femto)-BS	R_M (R_F)
Effective data rate of macro(femto)-BS	$\tilde{R}_M$ ($\tilde{R}_F/\tilde{R}_F^{(j)}$)
Path loss exponent	α
Noise power spectral density	N_0
Density of interfering macro(femto)-BSs	λ_M (λ_F)
Average effective data rate with MRT(ZFBF)	R_{MRT}(R_{ZFBF})
Average effective data rate with MRT–ZFBF	$R_{\text{MRT–ZFBF}}$

one may simply assign a single slot. Our analysis in the next section can be easily extended to model any desirable fairness among users and the associated user scheduling in different time slots.

4. PERFORMANCE ANALYSIS

For simplicity, we assume that the distances between the co-located users and the N neighboring femto-BSs (the macro-BS) are the same, i.e., $d_{j,k} \triangleq d_F$ ($d_{0,k} \triangleq d_M$), see Fig. 1b for a pictorial representation. Also, we assume that the macro-BS, the femto-BSs, and the users are all equipped with a single antenna, and we consider quasi-static Rayleigh flat-fading channels with unit mean power. We denote the transmit power of a macro-BS and a femto-BS by P_M and P_F, and the data rates of a macro-BS and a femto-BS by R_M and R_F. The bandwidth of the frequency slot is denoted by W, the path loss exponent by α, and the noise power spectral density by N_0.

To simplify the notation in the following derivations, we also define the effective data rate as the data rate multiplied by the non-outage (transmission success) probability (i.e., the probability that the channel can support the data rate [25]). Thus, for a macro-BS the effective data rate equals $\tilde{R}_M \triangleq R_M \cdot \Pr(W\log(1+\frac{P_M S_{\Gamma(1)} d_M^{-\alpha}}{N_0 W}) > R_M)$ and for a femto-BS cluster with a diversity of order j it equals $\tilde{R}_F^{(j)} \triangleq R_F \cdot \Pr(W\log(1+\frac{P_F S_{\Gamma(j)} d_F^{-\alpha}}{N_0 W}) > R_F)$, where $S_{\Gamma(1)}$ is an exponential random variable with unit mean (Rayleigh fading), and $S_{\Gamma(j)}$ is the sum of j i.i.d. exponential random variables with unit mean due to the coherent combining of the signals from j femto-BSs. Note that for further simplicity, when $j=1$ we denote $\tilde{R}_F^{(j)}$ as $\tilde{R}_F$. Table 1 summarizes the main notation used in the paper.

In the following, convex optimization is used in the performance analysis of randomized caching and MRT, while integer optimization is used in the performance analysis of threshold-based caching and ZFBF. Also, in Sections 4.1-4.4, we consider background noise without co-channel interference, and Section 4.5 extends the results in the presence of co-channel interference.

4.1 Randomized caching and MRT

We aim to maximize the average effective data rate of a typical user (denoted as R_{MRT}) with respect to the randomized caching parameters subject to the cache size constraint. We have the following optimization problem:

$$\max_{q_1,\cdots,q_M} \sum_{i=1}^{M} p_i U(q_i) \triangleq R_{\text{MRT}}$$
$$\text{subject to } 0 \le q_i \le 1,\ \forall i = 1,\cdots,M$$
$$\sum_{i=1}^{M} q_i \le m,$$

where p_i is the file popularity distribution (i.e., the probability that the typical user requests the ith file), q_i is the probability that the ith file is cached in a femto-BS and $U(q_i)$ is the effective data rate for the transmission of the ith file. (See Section 3.2.1 for a detailed description of the caching strategy.) Since we have N femto-BSs, the number of copies of the ith file in the femto-BSs is a binomial random variable with mean Nq_i. Thus, the number of cooperating femto-BSs for the transmission of the ith file (denoted as C_i) follows the law

$$\Pr(C_i = j) = \binom{N}{j} q_i^j (1-q_i)^{N-j},\ j = 0,1,\cdots,N.$$

The effective data rate for the transmission of the ith file, $U(q_i)$, can be computed as

$$U(q_i) = \Pr(C_i=0) R_M \Pr(\text{transmission success with macro-BS})$$
$$+ \sum_{j=1}^{N} \Pr(C_i = j) R_F \Pr(\text{trans. success with } j \text{ femto-BSs jointly})$$
$$= (1-q_i)^N \tilde{R}_M + \sum_{j=1}^{N} \binom{N}{j} q_i^j (1-q_i)^{N-j} \tilde{R}_F^{(j)}. \quad (1)$$

Since we have assumed Rayleigh fading, we can rewrite the effective data rates as $\tilde{R}_M = R_M e^{-\eta_M}$ and $\tilde{R}_F = R_F e^{-\eta_F}$ where $\eta_M = (2^{R_M/W}-1)N_0 W d_M^{\alpha}/P_M$ and $\eta_F = (2^{R_F/W}-1)N_0 W d_F^{\alpha}/P_F$. We have the following theorem.

THEOREM 1. *If $R_M e^{-\eta_M} \le R_F e^{-\eta_F}(1-\eta_F)$, then $U(q_i)$ is concave and thus $R_{MRT}(q_1,\cdots,q_M)$ is concave in the domain $0 \le q_i \le 1,\ i = 1,\cdots,M$.*

PROOF. The proof is provided in the Appendix. □

The condition $R_M e^{-\eta_M} \le R_F e^{-\eta_F}(1-\eta_F)$ implies that the marginal rate gain of including one more femto-BS into the cluster to perform cooperative transmission is decreasing and is smaller than the marginal gain of using a femto-BS instead of the macro-BS (see the appendix for a detailed discussion). These conditions usually hold in practice[4].

For the maximization problem with concave objective and linear constraints, we can find the optimal q_i^* by the KKT conditions. We have:

$$q_i^* = \left[U'^{-1}\left(\frac{\lambda}{p_i}\right)\right]_0^1,\ i = 1,\cdots,M;\ \sum_{i=1}^{M} q_i^* = m, \quad (2)$$

where λ is the Lagrange multiplier, $U'^{-1}(\cdot)$ is the inverse function of $U'(\cdot)$, which exists since $U'(\cdot)$ is monotonic (see Lemma 1 in the Appendix), and $[x]_0^1 \triangleq \min(\max(0,x),1)$. We also note that $U'^{-1}(\cdot)$ is a decreasing function. So, q_i^* increases as p_i increases, i.e., we have a higher probability to cache a more popular file, which makes sense intuitively.

[4]When these assumptions do not hold, one can still solve the optimization problem using non-convex optimization techniques.

4.2 Threshold-based caching and ZFBF

Here we aim to maximize the average effective data rate of the K users (denoted as $R_{\text{ZFBF}}(T)$), where T is the caching threshold in the threshold-based caching and ZFBF. To keep the exposition simple, we start by assuming that the number of users K is smaller than the number of femto-BSs N ($K \leq N$). We extend the analysis to the practical case with $K > N$ in the end of the section. Last, we define the cdf of the file popularity distribution as $v_j = \sum_{i=1}^{j} p_i,\ j = 1, \cdots, M$.

Since each of the K users generates a request in an i.i.d. manner, a user would be in the first group, the second group, and the third group with probability v_T, $v_{T+N(m-T)} - v_T$, and $1 - v_{T+N(m-T)}$, respectively. (See Section 3.2.2 for the definition of these three groups.) So, the probability that there are $K_1 = i$ users in the first group, $K_2 = j$ users in the second group, and $K_3 = K - K_1 - K_2 = K - i - j$ users in the third group is $\frac{K!}{i!j!(K-i-j)!} v_T^i (v_{T+N(m-T)} - v_T)^j (1 - v_{T+N(m-T)})^{K-i-j}$.

As shown in Fig. 3b, we associate the K users with K different time slots. The requested file of a user of the third group (with group size $K_3 = K - i - j$) only appears in the macro-BS and it is served by the macro-BS with rate $\tilde{R}_M$ in this user's dedicated time slot. The resulting group sum rate is $(K - i - j)\tilde{R}_M$. For a user of the second group (with group size $K_2 = j$), its requested file only appears in a single femto-BS and it is served by that femto-BS with rate $\tilde{R}_F$ in this user's dedicated time slot. Thus, the resulting group sum rate is $j\tilde{R}_F$. For each user of the first group (with group size $K_1 = i$), its requested file appears in all N femto-BSs. By using ZFBF we create i spatially isolated channels (i.e., all i users can be served simultaneously with rate $\tilde{R}_F$ without seeing interference from each other), achieving a multiplexing gain of order i. Simultaneous transmissions can occur in their i dedicated time slots, so each user of the first group achieves an i-fold increase in rate, $i\tilde{R}_F$. The resulting group sum rate is $i \cdot i\tilde{R}_F$.

The average effective data rate $R_{\text{ZFBF}}(T)$ of the K users in the three groups can be computed as

$$\begin{aligned} R_{\text{ZFBF}}(T) &= \frac{1}{K}\left\{\sum_{i=0}^{K}\sum_{j=0}^{K-i} \frac{K!}{i!j!(K-i-j)!}\right. \\ &\cdot v_T^i (v_{T+N(m-T)} - v_T)^j (1 - v_{T+N(m-T)})^{K-i-j} \\ &\left.\cdot \left[i \cdot i\tilde{R}_F + j\tilde{R}_F + (K-i-j)\tilde{R}_M\right]\right\} \\ &= \frac{1}{K}\left\{(\mathbb{E}[K_1^2] + \mathbb{E}[K_2])\tilde{R}_F + \mathbb{E}[K_3]\tilde{R}_M\right\} \qquad (3) \\ &= [v_{T+N(m-T)} + (K-1)v_T^2]\tilde{R}_F + (1 - v_{T+N(m-T)})\tilde{R}_M. \end{aligned}$$

Note that with the averaging $\frac{1}{K}\{\cdot\}$ in R_{ZFBF}, both R_{ZFBF} and R_{MRT} are consistently defined with respect to a unit time slot. We can find the optimal caching threshold T^* by enumeration of the solution space $T = 0, 1, \cdots, m$, which is linear in the cache size m and thus easy to compute in practice. Fig. 6 in Section 5 shows the resulting data rate under various threshold values for practical scenarios.

Example: We consider two special cases.

- When $T = m$, we cache the most popular m files in all N femto-BSs and we have $R_{\text{ZFBF}}(m) = [v_m + (K-1)v_m^2]\tilde{R}_F + (1 - v_m)\tilde{R}_M$. If $v_m \approx 1$, that is, we almost always request one of the m most popular files, we have $R_{\text{ZFBF}}(m) \approx K\tilde{R}_F$.
- When $T = 0$, we cache only one copy of the most popular Nm files in femto-BSs and we have $R_{\text{ZFBF}}(0) = v_{Nm}\tilde{R}_F + (1 - v_{Nm})\tilde{R}_M$. If $v_{Nm} \approx 1$, we have $R_{\text{ZFBF}}(0) \approx \tilde{R}_F$.
- We can see that for very skewed popularity distribution satisfying $v_m \approx 1$, the rate R_{ZFBF} with threshold $T = m$ is K times higher than that with threshold $T = 0$, where K is the maximum multiplexing gain.

We now discuss how to extend the above results for the case where the number of users K is larger than the number of femto-BSs N. When $K > N$ and there are $K_1 = i > N$ users in the first group, the maximum multiplexing gain that can be achieved in a time slot is N, limited by the number of neighboring femto-BSs. It is easy to see that with time sharing between $\binom{i}{N}$ subsets of users of the first group (i.e., we arbitrarily choose N users out of i users to form a subset and a time slot is divided for analytical purposes into $\binom{i}{N}$ sub-slots to time share between the $\binom{i}{N}$ subsets), each user would be served with rate $\binom{i-1}{N-1}\tilde{R}_F / \binom{i}{N}$ in a time slot (since a user belongs to exactly $\binom{i-1}{N-1}$ subsets). Simultaneous transmissions with ZFBF can occur in total $K_1 = i$ dedicated time slots, so each user of the first group achieves the rate $i\binom{i-1}{N-1}\tilde{R}_F / \binom{i}{N} = N\tilde{R}_F$. The resulting group sum rate is $i \cdot N\tilde{R}_F$. The average effective data rate can be written as

$$\begin{aligned} R_{\text{ZFBF}}(T) = \frac{1}{K}\sum_{i=0}^{K}\sum_{j=0}^{K-i} &\frac{K!}{i!j!(K-i-j)!} \\ &\cdot v_T^i (v_{T+N(m-T)} - v_T)^j (1 - v_{T+N(m-T)})^{K-i-j} \\ &\cdot \left[i \cdot \min\{i, N\}\tilde{R}_F + j\tilde{R}_F + (K-i-j)\tilde{R}_M\right]. \end{aligned}$$

Remark: It is evident that there is a tradeoff associated with the value of the design parameter T. When T is large, we benefit from the multiplexing gain but more redundant files are held in the caches, resulting in an increasing amount of requests towards the low-rate macro-BS (cache misses). When T is small, we lose the multiplexing gain but most of the files are in the caches of the femto-BSs, generating fewer requests towards the macro-BS. The optimal choice of T depends on the file popularity distribution.

4.3 Multiple thresholds

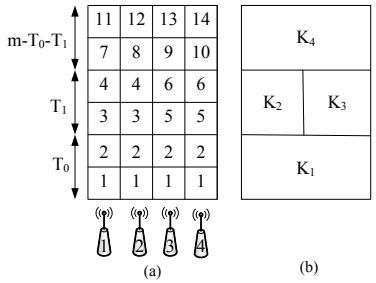


Figure 4: ZFBF and threshold-based caching with multiple thresholds ($m = 6,\ T_0 = 2,\ T_1 = 2$).

The above threshold-based caching scheme can be generalized to one with multiple thresholds where the basic idea is that the more popular a file is, the larger the number of copies of the file in the caches. Without loss of generality,

assume that the number of neighboring femto-BSs is $N = 2^n$ and define n thresholds $T_0, T_1, \cdots, T_{n-1}$, which are design parameters. The T_0 most popular files will be stored in all 2^n femto-BSs like before (2^n copies each), the next $2T_1$ most popular files will have 2^{n-1} copies each, and in general, the threshold T_i means that we allocate $2^n T_i$ storage units to cache $2^i T_i$ files, each with 2^{n-i} copies. Fig. 4a shows an example with $n = 2$ ($N = 4$).

In addition, the K users are divided into $2^n + 1$ groups according to their requests, where the size of the jth group is denoted as $K_j,\ j = 1, \cdots, 2^n$ (see Fig. 4b) and $K_{2^n+1} = K - K_1 - \cdots - K_{2^n}$ (the 2^n+1st group consists of users whose requested files only appear in the macro-BS). For example, in Fig. 4, if a user requests file "5", then the user belongs to group "3". Note that above we have defined more groups than thresholds, because different femto-BSs will store different files having the same number of copies, and thus they cannot simultaneously transmit any such set of files. For example, among files with two copies, only the sets (file "3", file "4") and (file "5", file "6") can be simultaneously transmitted, the former by femto-BSs "1" and "2" and the later by femto-BSs "3" and "4", thus the need to distinguish users in group "2" from those in group "3".

The probability that there are $K_j = k_j,\ j = 1, \cdots, 2^n+1$, users in the jth group is

$$\Pr(K_1 = k_1, \cdots, K_{2^n+1} = k_{2^n+1} = K - k_1 - \cdots - k_{2^n}) = \frac{K!}{k_1! k_2! \cdots k_{2^n+1}!} \prod_{l=0}^{n-1} \prod_{j=0}^{2^l-1} a_{2^l+j}^{k_{2^l+j}} a_{2^n}^{k_{2^n}} a_{2^n+1}^{k_{2^n+1}},$$

where a_g is the probability that a user belongs to the g-th group. For example, a_{2^l+j} is the probability that a user belongs to the $2^l + j$-th group, i.e., the user requests a file in the range of $\sum_{t=0}^{l-1} 2^t T_t + jT_l + 1$ to $\sum_{t=0}^{l-1} 2^t T_t + (j+1)T_l$, and thus $a_{2^l+j} = \sum_{i=\sum_{t=0}^{l-1} 2^t T_t + jT_l + 1}^{\sum_{t=0}^{l-1} 2^t T_t + (j+1)T_l} p_i$. Similarly, $a_{2^n} = \sum_{i=\sum_{t=0}^{n-1} 2^t T_t + 1}^{\sum_{t=0}^{n-1} 2^t T_t + 2^n (m - \sum_{t=0}^{n-1} T_t)} p_i$, and $a_{2^n+1} = 1 - a_1 - \cdots - a_{2^n}$. As a numerical example, in Fig. 4 the probability that a user belongs to group "3" is simply $p_5 + p_6$.

The average effective data rate can be computed as

$$R_{\text{ZFBF}}^{\text{multi}}(T_0, \cdots, T_{n-1}) = \frac{1}{K} \sum_{k_1=0}^{K} \sum_{k_2=0}^{K-k_1} \cdots \sum_{k_{2^n}=0}^{K-k_1-\cdots-k_{2^n-1}} \cdot \Pr(K_1 = k_1, K_2 = k_2, \cdots, K_{2^n} = k_{2^n}, K_{2^n+1} = k_{2^n+1}) \cdot \left[\sum_{l=0}^{n-1} \sum_{j=0}^{2^l-1} k_{2^l+j} \min\{k_{2^l+j}, 2^{n-l}\} \tilde{R}_F + k_{2^n} \tilde{R}_F + k_{2^n+1} \tilde{R}_M \right],$$

where the multiplexing gain that can be achieved in a time slot for the users in the $2^l + j$-th group is $\min\{k_{2^l+j}, 2^{n-l}\}$. To maximize $R_{\text{ZFBF}}^{\text{multi}}$, the optimal caching thresholds $T_i^*, i = 0, \cdots, n-1$ can be found by enumeration of the solution space $0 \le T_i \le m,\ i = 0, \cdots, n-1,\ \sum_{i=0}^{n-1} T_i \le m$, which is of size m^n. While this complexity is exponential in n, it is not very large in practice since the number of caching thresholds is quite small ($n = \log_2 N$). For example, even for relatively large clusters with, say, 8 cooperating femto-BSs there are at most 3 thresholds.

4.4 Joint MRT–ZFBF

We can achieve both diversity and multiplexing gain with a careful design of the ZFBF precoding [26]. Suppose there are N transmitters and K receivers ($K \le N$). Then, the N transmitters can transmit K independent (non-interfering or spatially isolated) streams to the K receivers simultaneously, each with a diversity of order $N - K + 1$ [26]. We will refer to this scheme as the MRT–ZFBF scheme. Assuming the (single) threshold-based caching and following the analysis in Section 4.2, the average effective date rate (denoted as $R_{\text{MRT–ZFBF}}(T)$) can be computed as

$$R_{\text{MRT–ZFBF}}(T) = \frac{1}{K} \sum_{i=0}^{K} \sum_{j=0}^{K-i} \frac{K!}{i! j! (K-i-j)!} \cdot v_T^i (v_{T+N(m-T)} - v_T)^j (1 - v_{T+N(m-T)})^{K-i-j} \cdot \left[i \cdot \min\{i, N\} \tilde{R}_F^{(N - \min\{i,N\}+1)} + j\tilde{R}_F + (K-i-j)\tilde{R}_M \right]. \quad (4)$$

When the number of users in the first group $K_1 = i < N$, each user of the first group can achieve the rate $\tilde{R}_F^{(N-i+1)}$ in a time slot with a diversity of order $N - i + 1$.

4.5 Noise and co-channel interference

We extend the analytical model to take into account co-channel interference from other macro-BSs and other femto-BSs outside the cooperation cluster using the same frequency slot at the same time with the typical user under study (i.e., the same resource block is spatially reused). We can approximate the spatial distribution of the interfering macro-BSs (outside the circle centered at the typical user with radius d_M, denoted as $B(0, d_M)$) as a Poisson point process Φ with density λ_M [7]. The co-channel interference from the interfering macro-BSs is denoted as I_M, $I_M = \sum_{x \in \Phi \setminus B(0,d_M)} P_M S_{x,\Gamma(1)} D_x^{-\alpha}$, where $S_{x,\Gamma(1)}$ denotes the fading gain of the interfering link from the xth interfering macro-BS to the typical user, which is exponentially distributed with unit mean (Rayleigh fading) and D_x is the distance between the xth interfering macro-BS and the typical user.

Similarly, we can approximate the spatial distribution of the interfering femto-BSs (outside the circle centered at the typical user with radius d_F, denoted as $B(0, d_F)$) as a Poisson point process Ψ with density λ_F. The co-channel interference from the interfering femto-BSs is denoted as I_F, $I_F = \sum_{y \in \Psi \setminus B(0,d_F)} P_F S_{y,\Gamma(1)} D_y^{-\alpha}$, where $S_{y,\Gamma(1)}$ and D_y are similarly defined.

Following the derivations in [15], the success (non-outage) probability of the transmission between the typical user and its serving macro-BS can be computed as

$$\begin{aligned} &\Pr(W \log(1 + \frac{P_M S_{\Gamma(1)} d_M^{-\alpha}}{I_M + I_F + N_0 W}) > R_M) \\ &= \exp\left(-\tau_M N_0 W d_M^{\alpha} P_M^{-1}\right) \exp\left\{-\lambda_M \pi d_M^2 \right. \\ &\cdot \left. \left(\tau_M^{\delta} \mathbb{E}_S[S^{\delta} \gamma(1-\delta, \tau_M S)] - \mathbb{E}_S[1 - \exp(-\tau_M S)]\right)\right\} \\ &\cdot \exp\left\{-\lambda_F \pi d_F^2 \left(\tau_M^{\delta} \xi^{\delta} \mathbb{E}_S[S^{\delta} \gamma(1-\delta, \tau_M \xi S)]\right.\right. \\ &\left.\left. -\mathbb{E}_S[1 - \exp(-\tau_M \xi S)]\right)\right\}, \end{aligned}$$

where $\tau_M = 2^{R_M/W} - 1$, $\delta = 2/\alpha$, $\xi = (\frac{d_M}{d_F})^{\alpha} \frac{P_F}{P_M}$, S is an exponential random variable with unit mean, and $\gamma(a, z) = \int_0^z \exp(-t) t^{a-1} dt$ is the lower incomplete gamma function.

In addition, the success probability of the transmission between the typical user and a cluster of j femto-BSs is

$$\Pr(W\log(1+\frac{P_F S_{\Gamma(j)} d_F^{-\alpha}}{I_M+I_F+N_0W})>R_F)=\sum_{k=0}^{j-1}\frac{1}{k!}(-1)^k\frac{d^k}{dt^k}V(t)\Big|_{t=1}$$

where $V(t)$ is

$$V(t)=\exp\left(-\tau_F N_0 W d_F^{\alpha} t P_F^{-1}\right)\exp\left\{-\lambda_F\pi d_F^2 \cdot\left(\tau_F^{\delta}t^{\delta}\mathbb{E}_S[S^{\delta}\gamma(1-\delta,\tau_F tS)]-\mathbb{E}_S[1-\exp(-\tau_F tS)]\right)\right\}$$
$$\cdot\exp\left\{-\lambda_M\pi d_M^2\left(\tau_F^{\delta}t^{\delta}\xi^{-\delta}\mathbb{E}_S[S^{\delta}\gamma(1-\delta,\tau_F t\xi^{-1}S)]-\mathbb{E}_S[1-\exp(-\tau_F t\xi^{-1}S)]\right)\right\}$$

and $\tau_F=2^{R_F/W}-1$. With these expressions, we can compute the effective data rates following the same steps as we did in Sections 4.1-4.4.

5. NUMERICAL RESULTS

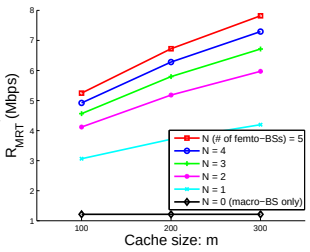
In this section we present performance results by numerically solving our analytical model in a number of practical scenarios. We assume that there is a library of $M=1000$ video files, and the file popularity distribution follows the Zipf distribution with parameter s, i.e., $p_i=c_{M,s}/i^s$, $i=1,\ldots,M$, where $c_{M,s}$ is the so called normalization constant.

Without loss of generality, consider a macro-BS with transmission range 4000m and a number of femto-BSs each with transmission range 200m deployed inside the macro-BS cell. Note that these are typical transmission ranges for a macro cell and for low-power BSs, see, for example, the capabilities of picocells, metrocells, and microcells defined in [1]. The number of femto-BSs inside a cluster (N), that is, the number of femto-BSs that a typical user can receive useful signal from, depends on the density of the femto-BSs and varies in the scenarios that we study. The data rate of the macro-BS is assumed to be $R_M=2$ Mbps and the data rate of the femto-BSs is assumed to be $R_F=10$ Mbps, again in line with industry practice. The transmit power of the macro-BS equals $P_M=20$ W and of the femto-BSs equals $P_F=20$ mW, as has been assumed in prior works as well [6].

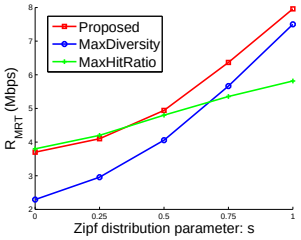
Without loss of generality, consider a cluster of femto-BSs covering an area of radius 200m at distance 2000m from the macro-BS. Then, the distance of the typical user from a femto-BS of the cluster is assumed to lie between 0 and 200m, and the distance from the macro-BS lies between 1800 and 2200m. We assume that the path loss exponent equals $\alpha=4$ and consider quasi-static Rayleigh flat-fading channels with unit mean power. In addition, the bandwidth of the frequency slot is $W=5$ MHz and the noise power spectral density varies from $N_0=4\times10^{-19}$ W/Hz to $N_0=8\times10^{-20}$ W/Hz. As a result, by substituting these numerical values into the formulas of Section 4, the transmission success (non-outage) probability for the macro- and the femto-BSs with a unit diversity varies from 0.6 to 0.9, and the effective data rate of the macro- and femto-BSs with a unit diversity varies from $\tilde{R}_M=2\times0.6=1.2$ to 1.8 Mbps and from $\tilde{R}_F=10\times0.6=6$ to 9 Mbps, respectively.

5.1 Data rates under diversity gains

We study the rate of a user under MRT achieved at an arbitrary time slot. To highlight the effect of diversity gains to the data rate, we show results when the success (non-outage) probability with a unit diversity equals 0.6. We later show results when the outage probability is smaller.



(a) Rate as a function of the cache size m for a cluster of N femto-BSs, $s=0.56$.



(b) Rate for three different caching approaches for different degrees of popularity skewness, $N=5$, $m=100$.

Figure 5: Performance of the system under MRT.

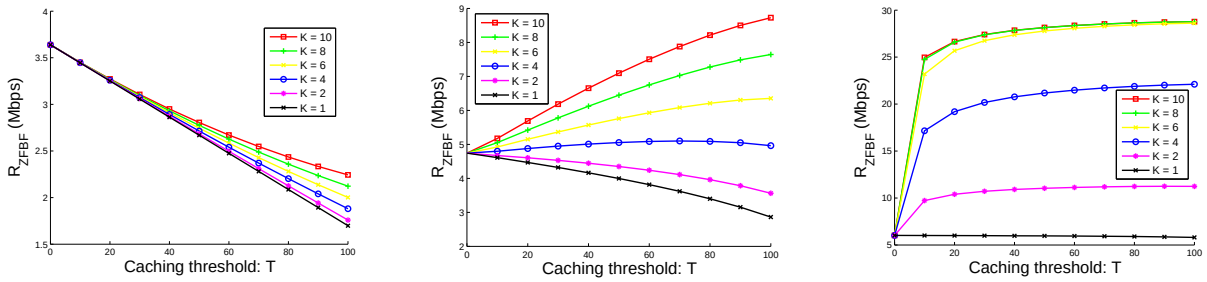
In Fig. 5a we plot the rate as a function of the cache size (m) and the number of nearby femto-BSs (N) for a popularity distribution with a typical parameter $s=0.56$ [24, 27]. As expected, the rate R_{MRT} increases with the cache size and the number of neighboring femto-BSs, the later because as N increases we have a larger diversity that reduces the transmission link failure (outage) probability. In this plot we also show the achieved rate when no femto-BSs are used. It is evident that using femto-BSs improves rates by 2-3x and adding MRT results in an additional gain of 2-3x. Last, note that for a fixed cache size and as we increase the number of femto-BSs, the marginal gain decreases since the outage probability has already been reduced to a very small value.

In Fig. 5b we plot the rate as a function of the Zipf distribution parameter s. We vary s from 0 (uniform distribution) to 1 (quite skewed distribution) and compare our scheme with the following two basic schemes: a "MaxDiversity" caching scheme and a "MaxHitRatio" caching scheme, to investigate how well our system adapts to changing levels of popularity. In the "MaxDiversity" scheme, we cache the most popular files 1 to m in every femto-BS so that we have a diversity of order N for all these files. Intuitively, this scheme would perform well for a skewed distribution with a large s. On the other hand, in the "MaxHitRatio" scheme we cache in the femto-BSs the most popular files 1 to mN, each with a single copy. This scheme would perform well for a near-flat (uniform) popularity distribution with a small s. As shown in Fig. 5b, our proposed cross-layer optimization scheme adapts to the popularity distribution and controls the diversity gain (equivalently, the number of copies of a file in the caches of the femto-BSs) for each individual file, attending a good performance in the whole range of s.

5.2 Data rates under multiplexing gains

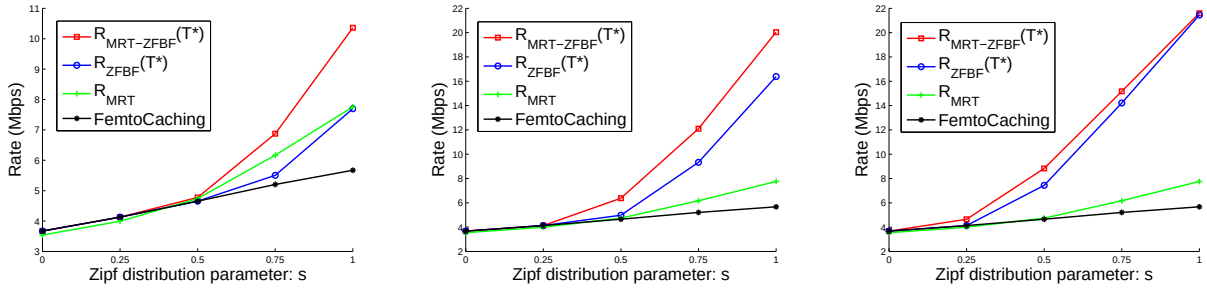
We study the rate under ZFBF achieved at an arbitrary time slot.

Fig. 6 plots the rate $R_{\text{ZFBF}}(T)$ as a function of the caching threshold T, $0\le T\le m$. We assume that the cluster of neighboring femto-BSs consists of $N=5$ femto-BSs, the cache size of each of those femto-BSs is $m=100$, and we consider three values for the Zipf distribution parameter, $s=0$, 0.56 and 1.5. In practice, a cluster of $N=5$ femto-BSs may serve tens to hundreds of users. In our study we are focusing on the users who receive transmissions from the macro- and femto-BSs at a single resource (time-frequency) block and thus vary their number from $K=1$ to 10.



(a) $s = 0$ (uniform popularity distribution) (b) $s = 0.56$ (moderately skewed) (c) $s = 1.5$ (very skewed)

Figure 6: Performance of the system under ZFBF, $N = 5$, $m = 100$.



(a) $K = 2$ (b) $K = 5$ (c) $K = 10$

Figure 7: Performance of the system under FemtoCaching only, MRT, ZFBF, and MRT–ZFBF, $N = 5$, $m = 100$.

From Fig. 6a we observe that for a uniform popularity distribution ($s = 0$) the optimal threshold is $T^* = 0$, i.e., we prefer to have only a single copy of a file in the caches to reduce cache misses and choose not to have a multiplexing gain. On the other hand, in Fig. 6c, for the very skewed popularity distribution ($s = 1.5$), the optimal threshold is $T^* = m = 100$ (for $K > 2$), i.e., we prefer to have multiple copies of the most popular files in the caches of femto-BSs, achieving a larger multiplexing gain. Note that for the very skewed distribution, we have $v_m \approx 1$ (v_m is the probability a user requests one of the m most popular files) and $R_{\text{ZFBF}}(m) \approx \min\{K, N\}\tilde{R}_F$, where the multiplexing gain is $\min\{K, N\}$. For a moderately skewed distribution, as shown in Fig. 6b with $s = 0.56$, the optimal threshold lies somewhere in the middle and depends on K (e.g., for $K = 4$, $T^* = 71$). Thus, our caching scheme can adapt to the file popularity distribution by properly setting the value of the caching threshold. Last, note that the rate gain from using ZFBF on the femto-BSs can be very substantial, reaching 10-20x in the case of skewed distributions.

5.3 Data rates using joint MRT–ZFBF

Fig. 7a-c compare the rates for $K = 2$, 5, and 10 users with "FemtoCaching only", MRT (R_{MRT}), ZFBF ($R_{\text{ZFBF}}(T^*)$), and MRT–ZFBF ($R_{\text{MRT–ZFBF}}(T^*)$), where in the latter two cases the rates are evaluated at their optimal caching thresholds, respectively. In the "FemtoCaching only" scheme, we cache in the femto-BSs the most popular files 1 to mN, each with a single copy, and there is no cooperative transmission among the femto-BSs.

The MRT–ZFBF scheme outperforms the ZFBF scheme for smaller values of K because there is some probability that the number of users in the first group K_1 is less than the number of femto-BSs N reducing the multiplexing gain. As a matter of fact, for $K = 2$ ZFBF performs worse than even plain MRT. This is because MRT/MRT–ZFBF achieves a diversity of order $N - K_1 + 1 \geq 2$ for all the users in the first group (see Eq. (4)) increasing the transmission success (non-outage) probability from 0.6 to more than 0.9 (equality for a diversity of 2), which, in turn, increases the effective data rate. Of course, when the number of users is large (e.g. $K = 10$) and thus $K_1 \geq N$ with high probability, ZFBF performs almost as well as MRT–ZFBF.

The results above are generated when the success (non-outage) probability with a unit diversity equals 0.6. For a larger success probability, e.g. 0.9, MRT is expected to perform relatively worse, and the ZFBF scheme is expected to be very close to MRT–ZFBF. Fig. 8 confirms the above. As a matter of fact, since the caching strategy under MRT caches files in a probabilistic fashion, when the file popularity is not very skewed it sometimes caches multiple copies of not so popular files which results in worse performance than even plain FemtoCaching (because the loss from the lower hit ratio is larger than the small gain from the diversity gain).

5.4 Single threshold vs. multiple thresholds

Fig. 9 plots the rate with ZFBF when two thresholds are used ($R^{\text{multi}}_{\text{ZFBF}}(T_0^*, T_1^*)$), over the rate when one threshold is used ($R_{\text{ZFBF}}(T^*)$) in a setup with 4 femto-BSs ($N = 4$). We observe that when the number of users is small (e.g. $K = 2$), the caching scheme with multiple thresholds is beneficial in the regime of skewed popularity (large s) because to get the maximum multiplexing gain it is sufficient to store 2 copies of the same file, which is possible under the multiple thresholds scheme. On the other hand, when K is large (e.g. $K = 8$) multiple thresholds are beneficial in the regime of near-flat popularity (small s). This is because for large s both schemes achieve the maximum multiplexing gain, whereas for small s multiple thresholds provide a finer granularity for controlling the trade-off between cache hit ratio and spatial multiplexing. Nevertheless, we can see

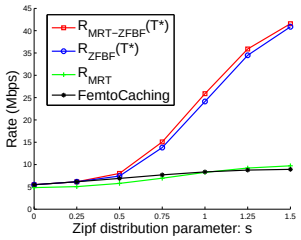


Figure 8: Performance of the system with a low outage probability of 0.1, $K = 5$, $N = 5$, $m = 100$.

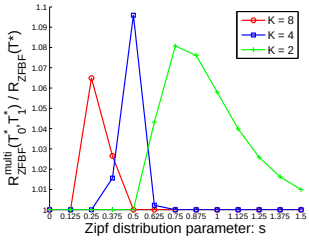


Figure 9: Performance of the system under ZFBF with single and multi-thresholds, $N = 4$, $m = 100$.

that $R^{\text{multi}}_{\text{ZFBF}}(T_0^*, T_1^*)$ is at most 10% larger than $R_{\text{ZFBF}}(T^*)$ across the whole range of file popularity distributions (s). Thus, the use of a single threshold appears to be enough to get most of the rate gains.

6. PRACTICAL CONSIDERATIONS

Updating the parameters of the caching strategy: The parameters of the caching strategy depend on the file popularity distribution p_i, e.g. q_i in Eq. (2) for MRT and the threshold T in Eq. (3) for ZFBF. The file popularity distribution can be estimated by user requests, see, for example, [8]. The complexity of computing the caching strategy parameter q_i in randomized caching under MRT is low since q_i's can be obtained efficiently by solving a convex optimization problem, see Theorem 1. The complexity of computing T in threshold-based caching is also tractable since the size of the solution space is linear in the cache size for the single threshold case. For the multi-threshold case the solution space is exponential in the number of thresholds, but (i) even for large cluster sizes, e.g. 8 femto-BSs, the number of thresholds is as small as 3, and (ii) as we have shown in Section 5.4 the use of a single threshold is enough to get most of the rate gains thus we do not anticipate the use of multiple thresholds in the majority of cases. The right place to perform the above computations is the macro-BS as it can keep track of all file requests and thus of the file popularity distribution, and it has more than enough computation power to compute the caching strategy parameters. Upon computing the parameters, the macro-BS will send their values to the femto-BSs. Last, since the time scale of significant changes in the file popularity distribution is in the order of a day or longer, updating the caching strategy parameters will happen infrequently.

Cache content update: Macro-BS and femto-BSs coordinate the cache content update (downloading popular video files via backhaul into caches in femto-BSs) according to updates in the caching strategy parameters. This can be done at off-peak hours because the time scale of significant changes in the file popularity distribution (e.g. days) is much larger than the time scale of receiving users' requests (e.g. seconds) [13]. Note also that only significant changes in the file popularity will result in a cache content update, while small changes in the file popularity will only result in reordering (relabeling) files in the caches.

The effect of asynchronous user requests: In the presence of many femto-BSs large rates occur when many users are served concurrently taking advantage of multiplexing gains. In practice this would occur when the system is highly loaded, which is the most relevant case since it is under high loads that it is a challenge to provide high rates. Under medium/low loads, the asynchronous nature of user requests poses the dilemma to wait till enough requests are collected to concurrently serve many users (which may result in extra delays) or to immediately serve less users resulting in a smaller multiplexing gain. Specifically, to get a multiplexing gain of i under threshold-based caching and ZFBF (see Section 4.2) there has to be i users that request anyone of the T most popular files "right now". While it is part of future work to fully investigate the effect of the asynchronous nature of user requests to performance, simply collecting requests for these most popular files while serving the users in the second and third groups (during time slots $1, \cdots, K_2$ and $1, \cdots, K_3$ in Fig. 3b), yields enough requests/users in the first group to achieve very large multiplexing gains (during the subsequent $1, \cdots, K_1$ time slots) without having to wait for any more requests. For typical system setup parameters, this achieved multiplexing gain is within 10% or closer to the maximum possible.

7. CONCLUSION

In this paper we proposed to jointly use and optimize distributed caching in femto-BSs and femto-BS cooperative transmissions. Our analytical and simulation results show that our system achieves an order of magnitude faster content delivery than legacy systems. The gains are particularly pronounced for skewed popularity distributions where caching multiple copies of popular files across multiple femto-BSs yields particularly large diversity and multiplexing gains without sizeably increasing cache misses. Given that content popularity is well known to be heavily skewed, our approach is expected to have a large impact in real-world setups.

8. REFERENCES

[1] http://www.smallcellforum.org/.

[2] H. V. Balan, R. Rogalin, A. Michaloliakos, K. Psounis, and G. Caire. Achieving high data rates in a distributed MIMO system. In *Proc. ACM MOBICOM*, Aug. 2012.

[3] H. V. Balan, R. Rogalin, A. Michaloliakos, K. Psounis, and G. Caire. Airsync: Enabling distributed multiuser MIMO with full spatial multiplexing. *IEEE/ACM Transactions on Networking*, Dec. 2013.

[4] D. Ben Cheikh, J. M. Kelif, M. Coupechoux, and P. Godlewski. Analytical joint processing multi-point cooperation performance in Rayleigh fading. *IEEE Wireless Commun. Lett.*, Aug. 2012.

[5] S. Borst, V. Gupta, and A. Walid. Distributed caching algorithms for content distribution networks. In *Proc. IEEE INFOCOM*, 2010.

[6] V. Chandrasekhar, J. G. Andrews, and A. Gatherer. Femtocell networks: a survey. *IEEE Commun. Mag.*, 46(9):59–67, Sept. 2008.

[7] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews. Modeling and analysis of k-tier downlink heterogeneous cellular networks. *IEEE J. Sel. Areas Commun.*, 30(3):550–560, Apr. 2012.

[8] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: characterizing popularity growth of YouTube videos. In *Proc. ACM WSDM*, 2011.

[9] D. Gesbert, S. Hanly, H. Huang, S. Shitz, O. Simeone, and W. Yu. Multi-cell MIMO cooperative networks: A new look at interference. *IEEE J. Sel. Areas Commun.*, 28(9):1380–1408, Dec. 2010.

[10] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo, H. S. Dhillon, and T. D. Novlan. Heterogeneous cellular networks: From theory to practice. *IEEE Commun. Mag.*, 50(6):54–64, June 2012.

[11] N. Golrezaei, A. Dimakis, and A. Molisch. Scaling behavior for device-to-device communications with distributed caching. *IEEE Trans. Inf. Theory*, 2014.

[12] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire. Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution. *IEEE Commun. Mag.*, Apr. 2013.

[13] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire. Femtocaching: Wireless video content delivery through distributed caching helpers. In *Proc. IEEE INFOCOM*, Mar. 2012.

[14] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE J. Sel. Areas Commun.*, 27(7):1029–1046, Sept. 2009.

[15] M. Haenggi and R. K. Ganti. Interference in large wireless networks. *Found. Trends Netw.*, 2009.

[16] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, and V. Jungnickel. Coordinated multipoint: Concepts, performance, and field trial results. *IEEE Commun. Mag.*, Feb. 2011.

[17] M. Ji, G. Caire, and A. F. Molisch. Fundamental limits of distributed caching in D2D wireless networks. In *Proc. IEEE ITW*, 2013.

[18] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana. Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges. *IEEE Commun. Mag.*, Feb. 2012.

[19] A. Liu and V. K. N. Lau. Mixed-timescale precoding and cache control in cached MIMO interference network. *IEEE Trans. Signal Process.*, Dec. 2013.

[20] A. Liu and V. K. N. Lau. Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems. *IEEE Trans. Signal Process.*, Jan. 2014.

[21] M. A. Maddah-Ali and U. Niesen. Decentralized coded caching attains order-optimal memory-rate tradeoff. *IEEE/ACM Trans. Netw.*, 2014.

[22] M. A. Maddah-Ali and U. Niesen. Fundamental limits of caching. *IEEE Trans. Inf. Theory*, May 2014.

[23] H. S. Rahul, S. Kumar, and D. Katabi. JMB: Scaling wireless capacity with user demands. In *Proc. ACM SIGCOMM*, Aug. 2012.

[24] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire. Femtocaching: Wireless content delivery through distributed caching helpers. *IEEE Trans. Inf. Theory*, Dec. 2013.

[25] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[26] K.-K. Wong and Z. Pan. Array gain and diversity order of multiuser miso antenna systems. *Int. J. Wireless Inf. Networks*, 2008.

[27] M. Zink, K. Suh, Y. Gu, and J. Kurose. Characteristics of YouTube network traffic at a campus network - measurements, models, and implications. *Comput. Netw.*, Mar. 2009.

Appendix: Proof of Theorem 1

First, let us define the function

$$G(q) \triangleq \sum_{j=0}^{N} \binom{N}{j} q^j (1-q)^{N-j} w_j,\ 0 \le q \le 1, \qquad (5)$$

where $w_j \ge 0,\ \forall j = 0, \cdots, N$. We have the following lemma.

LEMMA 1. *If $w_j \ge w_{j-1},\ \forall j = 1, \cdots, N$, then $G(q)$ is a non-decreasing function. Furthermore, if $w_{j+1} - w_j \le w_j - w_{j-1},\ \forall j = 1, \cdots, N-1$, then $G(q)$ is concave.*

PROOF. Using basic analysis, the first derivative of $G(q)$ can be computed as

$$G'(q) = \sum_{j=1}^{N} \frac{N!}{(j-1)!(N-j)!} q^{j-1}(1-q)^{N-j}(w_j - w_{j-1}) \ge 0,$$

where for the inequality we use the conditions $w_j \ge w_{j-1}$, $\forall j = 1, \cdots, N$. So, $G(q)$ is non-decreasing.

Similarly, the second derivative of $G(q)$ is computed as

$$G''(q) = \sum_{j=1}^{N-1} \frac{N!}{(j-1)!(N-j-1)!} q^{j-1}(1-q)^{N-j-1} \cdot [(w_{j+1} - w_j) - (w_j - w_{j-1})] \le 0,$$

where for the inequality we use the conditions $w_{j+1} - w_j \le w_j - w_{j-1},\ \forall j = 1, \cdots, N-1$. So, $G(q)$ is concave. □

Let w_0 represent the effective data rate of a macro-BS ($w_0 = \tilde{R}_M$) and w_j represent the effective data rate of a femto-BS cluster with a diversity of order j ($w_j = \tilde{R}_F^{(j)}$), $j = 1, \cdots, N$. From (1) and (5), we have $U(q_i) = \sum_{j=0}^{N} \binom{N}{j} q_i^j (1-q_i)^{N-j} w_j = G(q_i)$. In addition, since the ccdf of $S_{\Gamma(j)}$ is $\overline{F}_{S_{\Gamma(j)}}(z) = \sum_{n=0}^{j-1} \frac{1}{n!} e^{-z} z^n$, we obtain

$$\begin{aligned} w_{j+1} - w_j &= R_F \sum_{n=0}^{j} \frac{1}{n!} e^{-\eta_F} \eta_F^n - R_F \sum_{n=0}^{j-1} \frac{1}{n!} e^{-\eta_F} \eta_F^n \\ &= R_F \frac{1}{j!} e^{-\eta_F} \eta_F^j,\ \forall j = 1, \cdots, N-1; \\ w_1 - w_0 &= R_F e^{-\eta_F} - R_M e^{-\eta_M}. \end{aligned}$$

We note that the conditions $w_{j+1} \ge w_j,\ j \ge 0$ are equivalent to $R_F e^{-\eta_F} \ge R_M e^{-\eta_M}$. Also, the conditions $w_{j+1} - w_j \le w_j - w_{j-1},\ j \ge 2$ (the marginal rate gain of including one more femto-BS into the cluster to perform cooperative transmission is decreasing) are equivalent to $\eta_F \le 2$. Moreover, the condition $w_2 - w_1 \le w_1 - w_0$ (the aforementioned marginal gain is smaller than the marginal gain of using femto-BS instead of macro-BS) is equivalent to $R_M e^{-\eta_M} \le R_F e^{-\eta_F}(1 - \eta_F)$. As a result, by Lemma 1 and combining the above conditions, we conclude that if $R_M e^{-\eta_M} \le R_F e^{-\eta_F}(1 - \eta_F)$ holds, $U(q_i)$ is concave and thus $R_{\text{MRT}}(q_1, \cdots, q_M) = \sum_{i=1}^{M} p_i U(q_i)$ is concave.