On the delay of geographical caching methods in two-tiered heterogeneous networks

Ejder Bastug, Marios Kountouris, Mehdi Bennis, Merouane Debbah

On the Delay of Geographical Caching Methods in Two-Tiered Heterogeneous Networks

Ejder Baştuğ[◇], Marios Kountouris[○], Mehdi Bennis[†], and Mérouane Debbah[◇,○]

[◇]Large Networks and Systems Group (LANEAS), CentraleSupélec,
Université Paris-Saclay, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France
[○]Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris, France
[†]Centre for Wireless Communications, University of Oulu, Finland
{ejder.bastug, merouane.debbah}@centralesupelec.fr, marios.kountouris@huawei.com, bennis@ee.oulu.fi

***Abstract*—We consider a hierarchical network that consists of mobile users, a two-tiered cellular network (namely small cells and macro cells) and central routers, each of which follows a Poisson point process (PPP). In this scenario, small cells with limited-capacity backhaul are able to cache content under a given set of randomized caching policies and storage constraints. Moreover, we consider three different content popularity models, namely *fixed* content popularity, *distance-dependent* and *load-dependent*, in order to model the spatio-temporal behavior of users' content request patterns. We derive expressions for the average delay of users assuming perfect knowledge of content popularity distributions and randomized caching policies. Although the trend of the average delay for all three content popularity models is essentially identical, our results show that the overall performance of cached-enabled heterogeneous networks can be substantially improved, especially under the load-dependent content popularity model.**

***Index Terms*—edge caching, Poisson point process, stochastic geometry, mobile wireless networks, 5G**

I. Introduction

Content caching in 5G heterogeneous wireless networks improves the system performance, and is of high importance in limited-backhaul scenarios [1]. Most existing literature for cache-enabled heterogeneous networks using stochastic geometry focuses on the characterization of key performance metrics neglecting the backhaul limitations and the spatio-temporal content popularity profiles [2]–[5]. In order to capture these aspects, we analyze in this paper the gains of caching in heterogeneous network deployment considering the average delay as a performance metric.

Consider a multi-tier heterogeneous network where base stations in each tier are deployed according to a homogeneous Poisson point process (PPP). More precisely, we model a heterogeneous network which consists of mobile terminals (users), cache-enabled small base stations (SBSs), macro base stations (MBSs) and central routers. In this network setting, a user may experience delays due to downlink transmissions, backhaul and caches. Supposing that SBSs are able to cache contents proactively, we derive expressions for the average delay of typical users when connected to either MBSs or SBSs.

This research has been supported by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering), the projects BESTCOM and 4GinVitro, the Academy of Finland CARMA project and TEKES grant (2364/31/2014).

Moreover, in order to capture the spatio-temporal content access patterns of users, we suppose *fixed* content popularity, *distance-dependent* and *load-dependent* content popularities. Assuming that the content popularity distribution is perfectly known at the small base stations, we explore three different caching policies based on content-popularity and randomization.

II. System Model

Topology: We consider a multi-tier heterogeneous network in the two-dimensional Euclidean plane $\mathbb{R}^2$, where nodes in each tier k are distributed according to a homogeneous PPP $\Phi_k = \{r_i^{(k)}\}_{i\in\mathbb{N}}$ of intensity λ_k, and $r_i^{(k)} \in \mathbb{R}^2$ represents the location of the i-th node at the k-th tier. The above network layout models a multi-tier heterogeneous network that consists of mobile terminals (users), SBSs, MBSs and central routers with densities $\lambda_{\rm ut} > \lambda_{\rm sc} > \lambda_{\rm mc} > \lambda_{\rm cr}$, respectively. A *typical* mobile user is assumed to be located at the Cartesian origin $(0,0)$ in order to derive the performance metrics of the heterogeneous network.

Signal Model: We shall consider that the MBSs and SBSs are transmitting in the same frequency band and hence interfering with each other. The transmit power is $P_{\rm mc}$ for each MBS and $P_{\rm sc}$ for each SBS, where we assume that $P_{\rm mc} > P_{\rm sc}$. For notational convenience, let us denote a base station (transmitter) by its position. The received power experienced at a typical user due to a transmitter x is given by $P_x h_x \ell(x)$, where P_x is the transmit power ($P_{\rm mc}$ or $P_{\rm sc}$), h_x corresponds to the fast fading power coefficient (square of the fading amplitude) of the channel between transmitter x and typical user, and $\ell(x) = \|x\|^{-\alpha}$ is the standard power law pathloss function with $\alpha > 2$. The channel fading power coefficients are independent and identically distributed (i.i.d.) exponential random variables (Rayleigh fading) with $\mathbb{E}[h_x] = 1$.

Since we assume that the network is interference-limited (i.e., the interference power dominates over the noise power), we simply consider the signal-to-interference ratio (SIR). For a typical user connected to a MBS located at x, the SIR is given as

$$\mathrm{SIR}_{\rm mc}(x) = \frac{P_{\rm mc} h_x \ell(x)}{I_{\rm mm} + I_{\rm sm}} \tag{1}$$

where $I_{\mathrm{mm}} = \sum_{y \in \Phi_{\mathrm{mc}} \setminus \{x\}} P_{\mathrm{mc}} h_y \ell(y)$ is the interference experienced from all MBSs except the serving MBS at x, and $I_{\mathrm{sm}} = \sum_{y \in \Phi_{\mathrm{sc}}} P_{\mathrm{sc}} h_y \ell(y)$ is the aggregate interference experienced from SBSs. For a typical user connected to a SBS located at x, the SIR is given as

$$\mathrm{SIR}_{\mathrm{sc}}(x) = \frac{P_{\mathrm{sc}} h_x \ell(x)}{I_{\mathrm{ss}} + I_{\mathrm{ms}}} \tag{2}$$

where $I_{\mathrm{ss}} = \sum_{y \in \Phi_{\mathrm{sc}} \setminus \{x\}} P_{\mathrm{sc}} h_y \ell(y)$ is the interference experienced from all SBSs except the serving SBS, and $I_{\mathrm{ms}} = \sum_{y \in \Phi_{\mathrm{mc}}} P_{\mathrm{mc}} h_y \ell(y)$ is the aggregate interference from MBSs. The target SIR in our system model is denoted by γ.

Connectivity and Backhaul: Mobile user terminals are associated with the closest base station, either SBS or MBS. As alluded to earlier, each MBS or SBS is also connected to its nearest central router. Each central router has a high-rate broadband Internet connection. The wired backhaul is used to provide this broadband connection to MBSs and SBSs via backaul links, such that users' requests can be satisfied. Supposing that a content request is generated by a user, the base station is then in charge of starting immediately its distribution. An illustration of this heterogeneous network under limited-capacity backhaul is illustrated in Fig. 1.

A. Caching Model

When a user has a content request, we assume that the request is drawn from the distribution f_{pop}, which is in decreasing order of content popularities. More formally, the content popularity distribution of a user is a right continuous and monotonically decreasing probability distribution function (PDF), given by [6]

$$f_{\mathrm{pop}}(f, \eta) = \begin{cases} (\eta - 1) f^{-\eta}, & f \geq 1, \\ 0, & f < 1, \end{cases} \tag{3}$$

where f indicates a point in the support of the corresponding content, and $\eta > 1$ parametrizes the steepness of the popularity distribution curve.

In fact, higher values of η results in steeper distribution, which in turn means that certain contents are highly popular than the rest of contents in $f_{\mathrm{pop}}(f, \eta)$. Conversely, lower values of η yield a more uniform distribution, which in turns say that almost all contents have similar popularities. The content popularity of a user may be evolving over time and space, influenced by the choice of other users, and can be partially known at the base stations. This is somewhat equivalent to say that the parameter η can take different values depending on the scenario. In our case, each base station *perfectly* observes the content popularities according to three different models as follows:

- *Fixed*: The content popularity is identical for all users, with fixed steepness factor of $\eta = \eta_0$. Therefore, all SBSs observe the same distribution given by $f_{\mathrm{pop}}(f, \eta_0)$.
- *Distance-dependent*: The users have different content popularity distributions, each of them having a distance-dependent steepness factor $\eta = r$, where r is the (random) distance between a user and its serving SBS. Therefore, we assume that each SBS observes on average a content popularity distribution given by $f_{\mathrm{pop}}(f, \bar{r})$, where $\bar{r}$ is the average distance between the SBS and its users. This model is used to mimic the behavior of content popularity based on the distance (i.e., flat distribution in short distances).
- *Load-dependent*: The content popularity of users is load-dependent on average, each of SBS having parameter $\eta = \lambda_{\mathrm{ut}}/\lambda_{\mathrm{sc}}$. Therefore, all SBSs observe the content popularity distribution given by $f_{\mathrm{pop}}(f, \lambda_{\mathrm{ut}}/\lambda_{\mathrm{sc}})$. This model is used to mimic the behavior of content popularity based on the load (i.e., steep distribution in heavy loads).

Note that the choice of such a continuous content distribution is in fact for ease of analysis. When practical issues or analytical tractability are not a priority, Zipf-like discrete power laws can also be considered for modeling [6]. Indeed, content access statistics in cache-enabled web proxies [7], or more relevantly in base stations [8], [9] are characterized by such discrete power laws (or arguably distributions).

For the (some of) caching policies described below, we shall assume that the content popularity distribution $f_{\mathrm{pop}}(f, \eta)$ is perfectly known at the base stations. Practically, in order to have partial knowledge of $f_{\mathrm{pop}}(f, \eta)$ for the caching policies, statistical estimation methods can be employed either at the base stations in a distributed manner or alternatively at the central routers, by using statistical tools from machine learning (see [10], [11] for relevant discussions).

Given $f_{\mathrm{pop}}(f, \eta)$, the content in the interval $[1, f_0)$ is the *cacheable* content and is called as *catalogue*, whereas the remaining part $[f_0, \infty]$ is considered as non-cacheable content (i.e., sensor data, voice streaming and online gaming). An interval $[f, f + \Delta f)$ in the support of $f_{\mathrm{pop}}(f, \eta)$ is dedicated to represent the probability of the f-th content. Each SBS has a storage capacity of S, thus it caches contents according to a given caching policy. Having such a request behavior described above and caching capabilities at the SBSs, we consider the following *offline* caching policies:

- *StdPop* [12]: The most popular content from the catalogue is stored in the cache of SBSs and requires $S_{\mathrm{p}} \geq 0$ amount of storage. We additionally assume that the track of content popularity in a SBS requires S_0 amount of storage, defined as a function of the number of contents in the catalogue and the type of algorithm employed for content popularity estimation, thereby it holds that $S = S_{\mathrm{p}} + S_0 \geq 0$.
- *UniRand* [13]: The S amount of contents are cached uniformly at random. Note that this policy is not aware of the content catalogue, therefore it does not require any memory to track the content popularity profile.
- *MixPop*: The S_{p} amount of storage is used to cache the most popular content deterministically. The storage

overhead is $S_0 \geq 0$ and again defined as a function of number of content and the employed algorithm. In addition, we suppose that $S_{\rm u} \geq 0$ amount of storage is used to cache content uniformly at random, thus $S = S_{\rm p} + S_0 + S_{\rm u} \geq 0$.

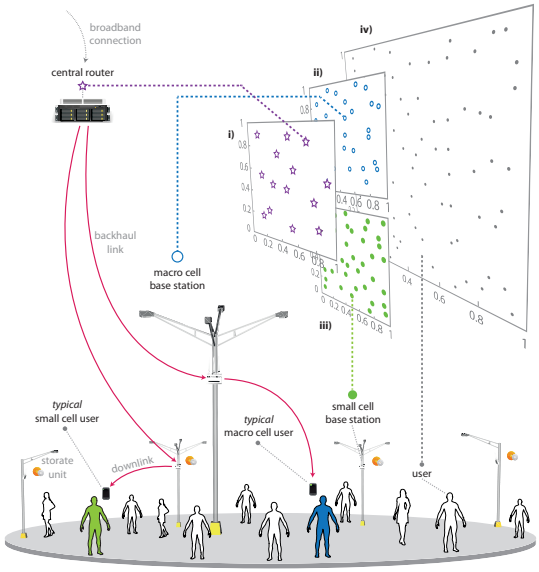

Figure 1: An illustration of the considered system model. The snapshots of i) central routers, ii) MBSs, iii) SBSs and iv) mobile user terminals are provided on the right side of figure.

In fact, if the catalogue size is sufficiently small, the storage overhead in StdPop and MixPop, due to the track of content popularity can be neglected. However, such an overhead may dominate the total storage space when a large catalogue with low-sized chunks is considered. One can also observe that the StdPop and UniRand policies are special cases of the MixPop policy and are given here for the sake of exposition.

The performance of any statistics-aware *online* cache removal policy (i.e., least-recently used (LRU) and least-frequently used (LFU)) would be upper bounded by its offline successor that has perfect content statistics; as such an online approach would require iterative estimation of content popularity in a finite time window, yielding to overall performance degradation. Such online policies can also be incorporated to our system model after some specific assumptions (see Independent Reference Model [14] for instance).

B. Delay and Quality of Service

Quality-of-service (QoS) is closely related to the delay experienced by users. We consider three different sources of delay which are detailed separately as follows.

Delay in downlink: When MBSs and SBSs have to deliver the contents to their intended mobile users, it is evident that the downlink transmission over the wireless medium incurs a delay mainly due to the interference from concurrent transmissions and channel fading. Consider now a simple retransmission protocol where a packet of requested content is repeatedly transmitted until its successful delivery, up to a pre-defined number of retransmission attempts M. Indeed, inferring whether a packet delivery is successful or not at the base station essentially relies on the signal-to-interference-plus-noise ratio (SINR) (or SIR in our case) being higher than the predefined threshold γ. If a packet is delivered successfully, we shall assume that the base station (macro or small cell) receives a one-bit acknowledgement message from the mobile user with negligible delay and error. Otherwise, if the delivery fails, the base station receives a one-bit negative acknowledgement message in the same vein. These attempts take T_0 amount of time. An *outage* event occurs if the packet is not delivered after M attempts. In the remainder, we denote the downlink delay experienced by the typical macro cell users (MUs) and small cell users (SUs) as $D_{\rm dm}$ and $D_{\rm ds}$ respectively.

Delay in backhaul: The delay caused in a wired backhaul link is modeled by an exponentially distributed random variable whose mean being proportional to the product of the average link distance (from typical base station to its nearest central router) and the average number of base stations connected to a single central router. In particular, representing the delay in macro and small cell backhaul links as $D_{\rm bm} \sim \text{Exponential}(\bar{\mu}_{\rm bm})$ and $D_{\rm bs} \sim \text{Exponential}(\bar{\mu}_{\rm bs})$ respectively, we (in general) suppose that $D_{\rm bs}$ stochastically dominates $D_{\rm bm}$.[1] This implies that small cell backhaul links are subject to higher delays compared to those of MBSs.

Delay in caches: Serving a user by fetching its content from the local cache is subject to delay as the storage medium is prone to errors, whereas such a delay may also vary depending on the storage type and the underlying mechanisms (i.e., hard disk, solid-state disk (SSD)). In this regard, we model this phenomenon as $D_{\rm ca} \sim \text{Exponential}(\bar{\mu}_{\rm ca})$, an exponentially distributed random variable with mean $\bar{\mu}_{\rm ca}$ being proportional to the storage type. We also assume that the delay of small cell backhaul links stochastically dominates the delay of reading a content from local caches, meaning that the speed of content reads from caches is stochastically higher than the speed of small cell backhaul links.

III. Performance Analysis

Considering the aforementioned sources of delay, namely downlink, caching and backhaul, the delay experienced by the typical MUs and SUs are respectively defined as

$$D_{\rm m} = D_{\rm dm} + D_{\rm bm}, \tag{4}$$

$$D_{\rm s} = D_{\rm ds} + \mathbb{1}_{\{f_{\rm s} \in \Delta_0\}} D_{\rm ca} + \big(1 - \mathbb{1}_{\{f_{\rm s} \in \Delta_0\}}\big) D_{\rm bs} \tag{5}$$

where $f_{\rm s}$ is the content requested by the typical small cell user and Δ_0 is the cache of its associated small cell. The indicator function $\mathbb{1}_{\{\dots\}}$ returns 1 if the statement holds, 0 otherwise. Before proceeding to the next step, let us define functions $B_1(T_0, M, \gamma, \alpha, P_x, P_y, \lambda_x, \lambda_y)$, $B_2(S_{\rm p}, \eta)$ and $B_3(S_{\rm u}, S_{\rm p}, f_0, \eta)$ given on the top of the next page.

[1] Given two random variables A and B, we say that A stochastically dominates B if $\mathbb{P}(A > x) \geq \mathbb{P}(B > x)$ for all a, or alternatively, $F_A(x) \leq F_B(x)$ for cumulative distribution functions $F_A(x)$ and $F_B(x)$.

$$B_1(T_0, M, \gamma, \alpha, P_x, P_y, \lambda_x, \lambda_y) = T_0 \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \frac{1}{1 + i\left[\rho(\gamma, \alpha) + (P_x/P_y)^{2/\alpha}(\lambda_x/\lambda_y)\gamma^{2/\alpha} A(\alpha)\right]} \tag{6}$$

$$B_2(S_{\mathrm{p}}, \eta) = 1 - \left(1 + S_{\mathrm{p}}\right)^{1-\eta} \tag{7}$$

$$B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \eta) = \frac{S_{\mathrm{u}}}{f_0 - S_{\mathrm{p}}}\Big(1 - \left(1 + f_0\right)^{1-\eta} + \left(1 + S_{\mathrm{p}}\right)^{1-\eta}\Big) \tag{8}$$

We now state the following result related to the average delay experienced by the typical MUs.

Theorem 1. *The average delay for a typical user connected to its nearest MBS is given by*

$$\bar{D}_{\mathrm{m}} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{sc}}, P_{\mathrm{mc}}, \lambda_{\mathrm{sc}}, \lambda_{\mathrm{mc}}) + \frac{1}{2}\beta\lambda_{\mathrm{mc}}\lambda_{\mathrm{cs}}^{-3/2} \tag{9}$$

where $B_1(T_0, M, \gamma, \alpha, P_{\mathrm{sc}}, P_{\mathrm{mc}}, \lambda_{\mathrm{sc}}, \lambda_{\mathrm{mc}})$ is given in (6)*. The parameter β is a scaling factor, relating to the importance of backhaul delay over the non-backhaul delay.*

Proof. See Appendix B.2 in [15]. □

In Theorem 1, the function B_1 models the average downlink delay whereas the remaining term in $\bar{D}_{\mathrm{m}}$ incorporates the average delay caused due to the backhaul. The summation of terms is due to the consideration of independent PPPs. We now turn our attention to SU with and without caching capabilities at the SBSs.

Corollary 1. *The average delay for a typical user connected to its nearest small cell (with no caching) is given by*

$$\bar{D}_{\mathrm{m}} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}}) + \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2} \tag{10}$$

where $B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}})$ is given in (6)*.*

Proof. The result is a direct application of Theorem 1, thus it is immediately proved by following similar steps given in Appendix B.2 in [15]. □

Theorem 2. *When* MixPop *caching policy is employed at the SBSs, the average delay for a typical user connected to its nearest small cell under fixed content popularity distribution is given by*

$$\bar{D}_{\mathrm{fix}}^{(\mathrm{mix})} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}}) + \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2} + \left(\bar{\mu}_{\mathrm{ca}} - \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2}\right)\Big(B_2(S_{\mathrm{p}}, \eta_0) + B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \eta_0)\Big) \tag{11}$$

where $B_2(S_{\mathrm{p}}, \eta_0)$ and $B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \eta_0)$ are given in (7) *and* (8) *respectively.*

In case of distance-dependent content popularity, the average delay is given by

$$\bar{D}_{\mathrm{dist}}^{(\mathrm{mix})} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}}) + \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2} + \left(\bar{\mu}_{\mathrm{ca}} - \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2}\right)\Big(B_2(S_{\mathrm{p}}, \frac{1}{2\sqrt{\lambda_{\mathrm{sc}}}}) + B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \frac{1}{2\sqrt{\lambda_{\mathrm{sc}}}})\Big). \tag{12}$$

In case of load-dependent content popularity, the average delay is given by

$$\bar{D}_{\mathrm{load}}^{(\mathrm{mix})} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}}) + \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2} + \left(\bar{\mu}_{\mathrm{ca}} - \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2}\right)\Big(B_2(S_{\mathrm{p}}, \frac{\lambda_{\mathrm{ut}}}{\lambda_{\mathrm{sc}}}) + B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \frac{\lambda_{\mathrm{ut}}}{\lambda_{\mathrm{sc}}})\Big). \tag{13}$$

Proof. See Appendix B.3 in [15]. □

The functions B_2 and B_3 in Theorem 2 are related to caching popular contents and caching uniformly at random respectively, and captures the cache hit behavior of the MixPop policy. By slightly modifying the steps in the proof of Theorem 2, similar results for StdPop and UniRand caching policies can be readily obtained. Note that the results above are based on the assumption that the typical users are connected to their nearest base stations.

In the above, we have provided the average delay expressions for typical MUs and SUs. The total average network delay, total network cost (including deployment and operational costs), and optimization of these metrics with respect to system design parameters are left for future work.

IV. Numerical Results

In this section, we numerically validate our approximations derived in the previous section. The impact of critical system parameters are discussed as follows.

Impact of MBS density λ_{mc}: The change of average delay with respect to the MBS density is given in Fig. 2a. Therein, as the number of MBSs increases, we observe an increment in average delay. This is mainly due to the backhaul as the delay in backhaul is proportional to the distance and average number of connected MBSs. In this setup, even though the average distance from a MBS to its central central router decreases (thus less delay in the backhaul), the increasing number of base stations contributes more to the average delay, thus yielding such a behaviour. On the other hand, the average delay in SBSs remains static in this setup. However, we note that the average delay experienced by a typical small cell user is reduced by adding caching capabilities at the base stations. For instance, when content popularity is load-dependent and caching policy is MixPop, the average delay is reasonably less than other candidates (including typical users with no caching at SBSs).

Impact of small cell density λ_{sc}: The change of the average delay with respect to the small cell density is depicted in Fig. 2b. Similarly to the previous figure for MBS density, we see that the average delay increases for all kind of small cell

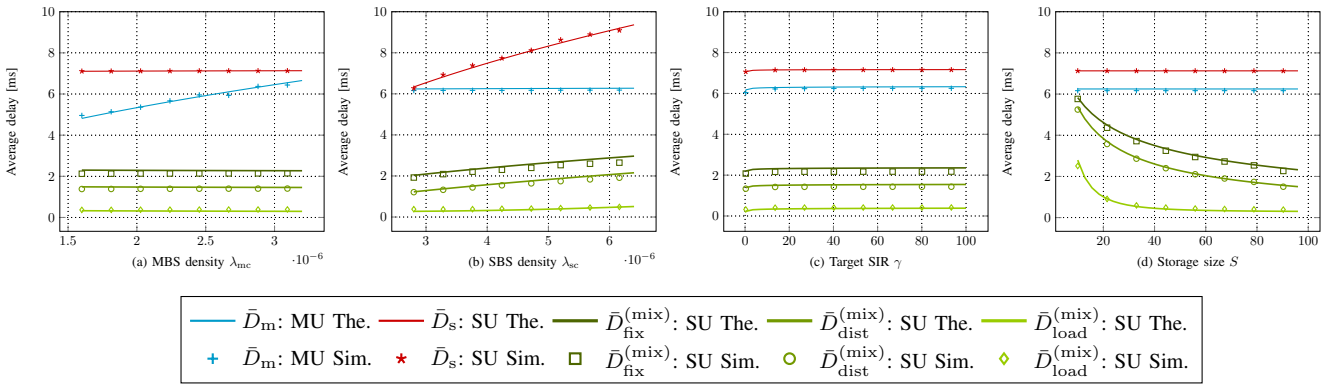

Figure 2: Evolution of average delay with respect to the a) macro cell density, b) small cell density, c) target SIR and d) storage size. $\lambda_{\text{cr}} = 1.4 \times 10^{-6}$, $\lambda_{\text{mc}} = 2.8 \times 10^{-6}$, $\lambda_{\text{sc}} = 3.6 \times 10^{-6}$, $\lambda_{\text{ut}} = 7.2 \times 10^{-6}$; $P_{\text{mc}} = 20$, $P_{\text{sc}} = 2$ Watts; $\alpha = 4$; $\gamma = 3$ dB; $M = 4$; $T_0 = 0.1$, $\mu_{\text{ca}} = 0.01$ ms; $\eta_0 = 1.45$; $f_0 = 500$, $S = 100$, $S_{\text{p}} = 9.5$, $S_0 = 0.5$, $S_{\text{u}} = 90$ GByte.

users. However, in this numerical setup, the rate of increment in delay with no-caching capabilities at the SBSs is higher than the delay experienced by the typical users with cache-enabled SBSs. Compared to the fixed and load-dependent content popularities, the typical user under load-dependent content popularity experiences less delay when the number of SBSs increases.

Impact of target SIR γ: In our setup, yet another important design parameter is the target SIR. In this regard, the average delay variation with respect to the target SIR is illustrated in Fig. 2c. As observed in the figure, the average delay increases by imposing higher target SIR values. This change is only visible in low values of target SIR, whereas the variation of delay in higher values of target SIR is negligible. This might stem from the fact that the downlink delay is not a dominating factor in our scenario compared to the backhaul delay. A typical user connected to the small cell with no caching capabilities experiences the highest delay, whereas the minimum delay is achieved by using MixPop policy under load-dependent content popularity. The delay of a typical MU remains between a SU with no-caching and caching capabilities at the base stations.

Impact of storage size S: Yet another crucial design parameter in our setup is the storage size. The impact of storage size on the average delay is shown in Fig. 2d. Indeed, as observed from the figure, dramatical decrease in delay is observed by increasing the storage size of small base stations. Similarly to previous observations, the most sensitive content popularity for the average delay is the load-dependent content popularity.

V. Conclusions

In this work, we have characterized the average delay of MUs and SUs under backhaul constraints and caching capabilities at the small base stations. Several content popularity distributions and caching policies have been considered. The main conclusion from this work is that caching at the small base stations allows for balancing the average access delay to the contents, especially if heterogeneous network densification under limited backhaul is considered.

References

[1] E. Baştuğ, M. Bennis, and M. Debbah, "Living on the Edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, August 2014.

[2] Z. Chen, J. Lee, T. Q. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *arXiv preprint arXiv:1601.00321*, 2016.

[3] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Modeling and performance analysis of clustered device-to-device networks," *arXiv preprint arXiv:1508.02668*, 2015.

[4] B. Serbetci and J. Goseling, "On optimal geographical caching in heterogeneous cellular networks," *arXiv preprint arXiv:1601.07322*, 2016.

[5] S. Yan, M. Peng, and W. Wang, "User access mode selection in fog computing based radio access networks," *arXiv preprint arXiv:1602.00766*, 2016.

[6] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[7] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *IEEE Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'99)*, vol. 1. IEEE, 1999, pp. 126–134.

[8] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*. ACM, 2011, pp. 305–316.

[9] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks," in *IEEE Conference on Computer Communications (INFOCOM)*, April 2015, pp. 936–944.

[10] G. Paschos, E. Baştuğ, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *arXiv preprint arXiv:1602.00173*, 2016.

[11] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Content-aware user clustering and caching in wireless small cell networks," in *11th International Symposium on Wireless Communication Systems (ISWCS'14)*, Barcelona, Spain, 2014.

[12] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking*, no. 1, p. 41, February 2015.

[13] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *IEEE International Conference on Communications (ICC)*, June 2015, pp. 3358–3363.

[14] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305–1314, September 2002.

[15] E. Baştuğ, "Distributed caching methods in small cell networks," Ph.D. dissertation, CentraleSupélec, Paris-Saclay University, December 2015. [Online]. Available: http://goo.gl/C22j1s