Delay and Stability Analysis of Caching in Heterogeneous Cellular Networks

Fatemeh Rezaei*, Babak H. Khalaj*, Ming Xiao[×], Mikael Skoglund[×]
*Electrical Engineering Department, Sharif University of Technology, Tehran, Iran
[×]Communication Theory Department, KTH, Stockholm, Sweden
Emails: f_rezaei@ee.sharif.edu, khalaj@sharif.edu, mingx@kth.se, skoglund@kth.se

Abstract—In this paper, we propose a general delay and stability performance analysis in Heterogeneous Cellular Caching Networks (HCCNs), based on queuing theory. We introduce new performance metrics in HCCNs and propose an optimization problem which minimizes the average experienced delay for users by ensuring the stability of the network. In addition, from the design perspective, we address the problem of finding the minimum cache size for the small cell base stations (SBSs) for having a tolerable average delay and also a stable network. Finally, the analytic expressions derived in this paper are validated through real trace-driven experiments on traffic of YouTube video requests.

Keywords—caching; delay; heterogeneous cellular networks; performance analysis; stability.

I. Introduction

Effective ways to reduce duplicate content transmissions by adopting intelligent caching strategies at the network level have been recently proposed [1]. Two key observations have encouraged a further use of caching in communication networks: (i) a large amount of traffic is due to a few number of popular files and (ii) the cost of disk storage has been reducing at a higher pace compared to many other components in communications/processing systems [2].

In this work, we provide a new performance analysis of caching in Heterogeneous Cellular Networks (HCNs), based on queuing theory. We consider a heterogeneous cellular network consisting of a single macro cell in which N small cell base stations (SBSs) operate in conjunction with the macro cell base station (MBS). The SBSs in such networks are equipped with cache memories. We call such networks as Heterogeneous Cellular Caching Networks (HCCNs). In this paper, the realistic assumption of stochastic arrival time and the traffic model of users requests is considered. We introduce new performance metrics in HCCNs and propose an optimization problem which minimizes the average experienced delay for users by ensuring the stability of the network. Related works studying delay minimization in HCCNs [2-4], provided combinatorial optimization problems with integer variables that were NP-hard and led to approximation algorithms. In contrast, in this paper, we propose a continuous variable optimization problem in HCCNs which can be handled through less complex solutions.

We propose a general delay and stability performance analysis in HCCNs for arbitrary caching schemes. In addition, from the design perspective, we address the problem of finding the minimum cache size for the SBSs in order to achieve a tolerable average delay and also a stable network.

The paper is organized as follows. Section II describes the system model. In Section III, the main analysis and results in HCCNs are proposed. In Section IV, the performance evaluation through trace driven experiment results is presented. Finally, Section V concludes the paper.

A. Related Works

Some recent works, [2-4], studied the role of caching in HCNs. The authors in [2] consider a cellular network model in which femto-basestations with low-rate backhaul equipped with considerable storage space assist the macro base station (BS). This work assumes that users can directly obtain files from the macro BS with the maximum delay, and subsequently formalizes the delay minimization problem, which is shown to be NP-hard.

Approximation algorithms for local caching of popular content items at small cell base stations have been proposed in [3], considering the bandwidth constraints of SBSs. The authors in [3] study approximation algorithms for the NP-hard problem of maximizing the fraction of content requests served locally by the deployed SBSs.

Finally, the authors in [4] design distributed caching optimization algorithms via belief propagation (BP) for minimizing the downloading latency in HCNs. In this work, a wireless interference channel is considered and the best SBS selection is used.

On the other hand, we studied single bottleneck caching networks from queuing theory perspective in [5] and provided an analysis of the stability, throughput, load on the bottleneck link and average response delay for various caching schemes in such networks.

II. System Model

We consider a heterogeneous cellular network consisting of a single macro cell in which N SBSs operate in conjunction with the MBS, as shown in Fig. 1. We

This work was supported in part by the Iran National Science Foundation under Grant 92017806.

978-1-5090-1990-8/16/$31.00 ©2016 IEEE

denote by s_0 the MBS and by $S = \{s_1, s_2, \dots, s_N\}$ the set of the SBSs where s_n, $n \in \boldsymbol{\mathcal{N}} = \{1,2,\dots,N\}$ represents the *n-th* SBS. The set of mobile users (MUs) submitting their content requests to the mobile network operator (MNO) is denoted by $M = \{m_1, m_2, \dots, m_U\}$, where m_u, $u \in \boldsymbol{\mathcal{U}} = \{1,2,\dots,U\}$ represents the *u-th* MU. Moreover, $M_n \subseteq M$, $n \in \boldsymbol{\mathcal{N}} = \{1,2,\dots,N\}$ represents the *n-th* users group such that the MUs in M_n are in the coverage area of the SBS s_n. The group M_0 also represents the users that only access the MBS without having access to any SBS. We assume that there is no overlap between the regions. That is, each user accesses only one SBS; extension of this work to the case of overlapping regions is beyond the scope of this paper and will be studied in future works.

We assume that the requests are drawn from a specific same-size file library $\mathbb{F} = \{f_i\}$, $i \in \boldsymbol{\mathcal{F}} = \{1,\dots,F\}$ of size B bits, and the cache content of SBS s_n, which is denoted by z_n, is a subset of the library $\mathbb{F}$. Each SBS s_n is capable of storing C whole files (i.e. *CB* bits).

The number of the content requests from the users connected to s_n within a second, i.e. the average request arrival rate at s_n, is denoted by λ_{req}^{n} [files per second].

Definition 1: We define $p_{req}(i,n)$ as the probability that file f_i is requested from the users connected to s_n within a transmission time slot of length τ.

We present the equations for general traffic of requests. In order to derive closed form equations, we also present results assuming that the stream of the requests conforms to the Independent Reference Model (IRM) traffic model, which is based on the following assumptions: i) users request files from a fixed library of F files; ii) the probability p_i that a request concerns file f_i, $i \in \boldsymbol{\mathcal{F}}$, is constant (*i.e.*, the file popularity does not vary over time) and is also *independent* of all past requests, generating an independent identically distributed (i.i.d.) sequence of requests.

Definition 2: The hit probability for the file f_i at s_n is denoted by $p_{hit}(i,n)$.

The average hit probability at SBS s_n is obtained from

$$p_{hit}(n) = \sum_{i=1}^{F} p_i p_{hit}(i,n)\,, \forall n \in \boldsymbol{\mathcal{N}}\,. \quad (1)$$

In such a network, the MUs submit their requests to the MNO. If the requested content for users in M_n is present in the cache of the SBS s_n, it is served locally by that SBS; otherwise, the content is sent to the user directly from the MBS. We assume that the MBS will support an average downlink rate denoted by r_0 [bps] for the MUs in the channels which are orthogonal to the channels spanning from the SBSs to the MUs with the average downlink rates denoted by r_n [bps]. We consider slotted transmission at the MBS and SBSs with a transmission time slot of length τ [sec].

Definition 3: The average response delay, $\bar{D}$, is defined as the average delay experienced by any given user in the HCCN, for obtaining the requested files, either delivered from the SBSs or from the MBS.

Since the channel spanning from each SBS to the MUs is shared among all users, there is a competition for each user to receive its requested file via the corresponding downlink path. On the other hand, the requests that are missed in the caches of the corresponding SBSs enter the MBS, where the channel from MBS to the users is also a shared bottleneck link. Therefore, we model the function of the MBS and SBSs downlinks by controlled FIFO queues where control units ensure that when multiple users request the same file concurrently (requests overlap within a transmission time slot), the MBS or SBSs only store the file in a single location of their queue.

The average requests arrival rate and service rate at the MBS and SBSs transmission queues are denoted by λ_n and μ_n for $n \in \boldsymbol{\mathcal{N}} \cup \{0\}$, respectively. The service time in our queue models have a general (arbitrary) distribution, with a mean and standard deviation, $\frac{1}{\mu_n}$ and σ_n, respectively. Therefore, we consider general G/G/1 queue models, where inter-arrival and service times have arbitrary distributions. In case of the IRM traffic model, the MBS and SBSs downlink functionalities are then modeled by M/G/1 queues.

By means of the proposed queue models, we introduce the MBS and SBSs utilization factors, i.e. $\rho_n \triangleq \frac{\lambda_n}{\mu_n}$, $n \in \boldsymbol{\mathcal{N}} \cup \{0\}$, as comprehensive performance metrics for the load on the corresponding links of heterogeneous cellular caching networks. The key advantage of such metrics is that they simultaneously take into account the cache hit probability, link loads, and requests arrival rates. Moreover, $\rho_n's$ provide key insight on the stability characteristic and bounded or unbounded delay behavior of such networks.

III. Main Analysis and Results

Some recent works have investigated the problem of delay minimization in HCCNs, as introduced in Section I. We are also interested in minimizing the user experienced delay as an important performance metric in such networks. In this paper, we propose a novel analysis based on queuing theory for this problem and provide a new perspective in such networks.

By definition, the average response delay for the users in M_n is obtained from

$$\overline{D_n} = p_{hit}(n)\,\overline{dl_n} + \left(1 - p_{hit}(n)\right)\overline{dl_0}\,, \quad (2)$$

where the average delay via the downlink path, $\overline{dl_n}$, $\forall n \in \boldsymbol{\mathcal{N}} \cup \{0\}$, is given by the sum of the average service time, i.e $\frac{1}{\mu_n}$, and the average time spent in the corresponding queue. The average response delay in the network is also given by

$$\bar{D} = \frac{1}{N}\sum_{n=1}^{N} \overline{D_n}\,. \quad (3)$$

In case of IRM traffic, we consider M/G/1 queue models. According to the Pollaczek-Khinchin (P-K) relation

[6], the average downlink delay is given by $\overline{dl_n} = \frac{1}{\mu_n}(1 + \frac{\rho_n(1+v_n^2)}{2(1-\rho_n)})$, $\forall n \in \boldsymbol{\mathcal{N}} \cup \{0\}$, where $v_n = \mu_n \sigma_n$ is the coefficient of variation of the service time. Therefore, in case of IRM traffic, $\overline{D_n}$ is obtained from

$$\overline{D_n} = p_{hit}(n)\, \frac{1}{\mu_n}(1 + \frac{\rho_n(1+v_n^2)}{2(1-\rho_n)}) + \big(1 - p_{hit}(n)\big)\frac{1}{\mu_0}(1 + \frac{\rho_0(1+v_0^2)}{2(1-\rho_0)}) \ . \tag{4}$$

Equation (4) shows that the average response delay of HCCNs in stable regions, i.e. $\rho_n < 1$, $\forall n \in \boldsymbol{\mathcal{N}} \cup \{0\}$, is bounded and is given as a function of the hit probabilities and utilization factors. In [5], we have presented the utilization factor of a content server connected through a shared link to a number of stations, where the stations are homogeneous and there is no constraint on their capacity, i.e. unlimited capacity between stations and users. In this paper, we consider the problem of delay minimization in HCCNs, considering the stability constraints in such networks. In HCCNs, according to the described model, the SBSs also have a constrained capacity. Therefore, in order to analyze such networks, we define and derive the corresponding SBS utilization factors.

Proposition 1: The utilization factor of the queue of SBS s_n, $n \in \boldsymbol{\mathcal{N}}$, in HCCNs is derived as

$$\rho_n = \frac{B}{r_n \tau} \sum_{i=1}^{F} p_{req}(i,n) p_{hit}(i,n) \ . \tag{5}$$

Proof: We define the random variable $X_{i,n}$ as the number of the requests for file f_i arriving at SBS s_n within a transmission time slot and being hit at the cache of s_n. In order to model the scenario where multiple users may request the same file at a given time slot, a control unit ensures that packets of the repeated requested files enter the SBS queue only once at each time slot. The function $h(x) = x + (1-x)u(x-1)$, where $u(x)$ denotes the step function is designed to provide such functionality. Consequently, the requests for the specific file f_i enter the SBS s_n queue with the average arrival rate $\lambda_{f_i}^n$ given by

$$\lambda_{f_i}^n = \mathbb{E}\big[h(X_{i,n})\big] = \sum_{k=0}^{\infty} h(k) \mathrm{P}\big(X_{i,n} = k\big) = \sum_{k=1}^{\infty} \mathrm{P}\big(X_{i,n} = k\big) = 1 - \mathrm{P}\big(X_{i,n} = 0\big) = p_{req}(i,n) p_{hit}(i,n) \ . \tag{6}$$

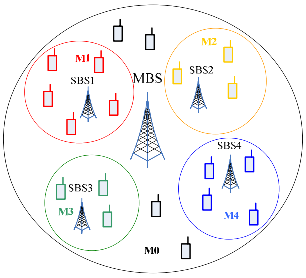

Fig. 1. System model.

Therefore, the average arrival rate at the SBS s_n queue, i.e. λ_n, is obtained from

$$\lambda_n = \sum_{i=1}^{F} \lambda_{f_i}^n = \sum_{i=1}^{F} p_{req}(i,n) p_{hit}(i,n) \ . \tag{7}$$

In addition, according to the SBS transmission capacity and file size constraints, the average service rate at the SBS queue is obtained from $\mu_n = \frac{r_n}{B}\tau$ [files per time slot]. Consequently, given $\rho_n \triangleq \frac{\lambda_n}{\mu_n}$, (5) is derived.■

Proposition 2: The utilization factor of the MBS queue in HCCNs is derived as

$$\rho_0 = \frac{B}{r_0 \tau} \sum_{i=1}^{F} \Big(1 - \prod_{n=1}^{N} \Big(1 - p_{req}(i,n)\big(1 - p_{hit}(i,n)\big)\Big) \Big). \tag{8}$$

Proof: The proof is similar to the approach in the proof of Theorem 1 in [5], without the assumption of homogeneous stations, and taking into account that the average service rate at the MBS queue in this case is given by $\mu_0 = \frac{r_0}{B}\tau$ [files per time slot]. ■

Lemma 1: In case of IRM traffic, $p_{req}(i,n)$ is obtained from

$$p_{req}(i,n) = 1 - e^{-\lambda_{req}^n p_i \tau} \ . \tag{9}$$

Proof: We define the random variable $N_{i,n}$, as the number of the requests for file f_i arriving at SBS s_n within a time slot. According to Definition 1, we have

$$p_{req}(i,n) = 1 - P\big(N_{i,n} = 0\big) \ . \tag{10}$$

In case of IRM traffic, the distribution of $N_{i,n}$ is Poisson with the mean $\lambda_{req}^n\, p_i \tau$, which results in (9). ■

Theorem 1: In HCCNs with IRM traffic for different caching schemes, the average response delay in stable regions is obtained from

$$\overline{D} = \frac{1}{N} \sum_{n=1}^{N} \left(p_{hit}(n) \frac{B}{r_n \tau} \left(1 + \frac{(1+v_n^2)\frac{B}{r_n \tau}\sum_{i=1}^{F} \left(1-e^{-\lambda_{req}^n p_i \tau}\right) p_{hit}(i,n)}{2\left(1-\frac{B}{r_n \tau}\sum_{i=1}^{F} \left(1-e^{-\lambda_{req}^n p_i \tau}\right) p_{hit}(i,n)\right)}\right) + \left(1 - p_{hit}(n)\right) \frac{B}{r_0 \tau} \left(1 + \frac{(1+v_0^2)\frac{B}{r_0 \tau}\sum_{i=1}^{F}\left(1-\prod_{n=1}^{N}\left(1-\left(1-e^{-\lambda_{req}^n p_i \tau}\right)\left(1-p_{hit}(i,n)\right)\right)\right)}{2\left(1-\frac{B}{r_0 \tau}\sum_{i=1}^{F}\left(1-\prod_{n=1}^{N}\left(1-\left(1-e^{-\lambda_{req}^n p_i \tau}\right)\left(1-p_{hit}(i,n)\right)\right)\right)\right)}\right)\right). \tag{11}$$

Proof: Combining (3-5) and (8-9) results in (11). ■

Theorem 1 provides the average response delay in HCCNs as a function of the cache hit probabilities and network parameters.

Therefore, in order to minimize the average response delay in HCCNs, based on Proposition 1-2 and Theorem 1, the optimization problem is formulated as:

$$\min \overline{D} = \frac{1}{N}\sum_{n=1}^{N}\Bigg(p_{hit}(n)\frac{B}{r_n\tau}\Bigg(1+ \frac{(1+v_n^2)\frac{B}{r_n\tau}\sum_{i=1}^{F}\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)p_{hit}(i,n)}{2\left(1-\frac{B}{r_n\tau}\sum_{i=1}^{F}\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)p_{hit}(i,n)\right)}\Bigg) + \Big(1- p_{hit}(n)\Big)\frac{B}{r_0\tau}\Bigg(1+\frac{(1+v_0^2)\frac{B}{r_0\tau}\sum_{i=1}^{F}\left(1-\prod_{n=1}^{N}\left(1-\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)(1-p_{hit}(i,n))\right)\right)}{2\left(1-\frac{B}{r_0\tau}\sum_{i=1}^{F}\left(1-\prod_{n=1}^{N}\left(1-\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)(1-p_{hit}(i,n))\right)\right)\right)}\Bigg)\Bigg) \tag{12}$$

$s.t.$

$$\rho_0 = \frac{B}{r_0\tau}\sum_{i=1}^{F}\left(1-\prod_{n=1}^{N}\left(1-\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)\left(1-p_{hit}(i,n)\right)\right)\right) < 1 \tag{13}$$

$$\rho_n = \frac{B}{r_n\tau}\sum_{i=1}^{F}\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)p_{hit}(i,n) < 1, \quad \forall n \in \boldsymbol{\mathcal{N}} \tag{14}$$

$$\sum_{i=1}^{F} p_{hit}(i,n) = |z_n| \qquad \forall n \in \boldsymbol{\mathcal{N}} \tag{15}$$

$$0 \le p_{hit}(i,n) \le 1 \qquad \forall i \in \boldsymbol{\mathcal{F}}, \forall n \in \boldsymbol{\mathcal{N}}\ , \tag{16}$$

where $p_{hit}(i,n),\ \forall i \in \boldsymbol{\mathcal{F}},\ \forall n \in \boldsymbol{\mathcal{N}}$ is an unknown variable. Equations (13) and (14) are stability constraints (it should be noted that (12) is only valid for stable regions), and (15) is the cache size constraint. This is a non-convex constrained optimization problem which we can approach based on approximation algorithms.

In order to reduce the complexity, we consider the problem from another perspective. By choosing the cache replacement policy, $p_{hit}(i,n)$ is known. By applying Theorem 1, and using the formulas of $p_{hit}(i,n)$ for different caching schemes provided in [5] and [7], the average response delay in terms of the network parameters is obtained.

From a design perspective, the goal is to find the minimum cache size such that the average response delay becomes less than the maximum tolerable delay, i.e. D_{max}, that is

$$C^* = \min\{C \le F : \overline{D} < D_{max}\}\ . \tag{17}$$

In [5], we have shown the effect of various caching replacement policies such as Least Recently Used (LRU), Least Frequently Used (LFU) and random (RAND), on the network performance in terms of the average response delay, hit probability, stability and utilization factor. In fact, LFU has shown the best performance among the aforementioned cache replacement policies. The average response delay for LFU is obtained by the following corollary.

Corollary 1: In HCCNs with IRM traffic for LFU caching schemes, the average response delay in stable regions is obtained from

$$\overline{D} = \frac{1}{N}\sum_{n=1}^{N}\Bigg(\left(\sum_{i=1}^{C} p_i\right)\frac{B}{r_n\tau}\Bigg(1+ \frac{(1+v_n^2)\frac{B}{r_n\tau}\sum_{i=1}^{C}\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)}{2\left(1-\frac{B}{r_n\tau}\sum_{i=1}^{C}\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)\right)}\Bigg) + \left(\sum_{i=C+1}^{F} p_i\right)\frac{B}{r_0\tau}\Bigg(1+\frac{(1+v_0^2)\frac{B}{r_0\tau}\sum_{i=C+1}^{F}\left(1-\prod_{n=1}^{N}\left(1-\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)\right)\right)}{2\left(1-\frac{B}{r_0\tau}\sum_{i=C+1}^{F}\left(1-\prod_{n=1}^{N}\left(1-\left(1-e^{-\lambda_{req}^{n}p_i\tau}\right)\right)\right)\right)}\Bigg)\Bigg). \tag{18}$$

Proof: The hit probability for file f_i in a cache with the LFU replacement policy is given by

$$p_{hit}(i,n) = \begin{Bmatrix} 1 & \forall i \in \{1,\dots,C\} \\ 0 & O.W. \end{Bmatrix},\ \ \forall i \in \{1,\dots,F\}\,. \tag{19}$$

Inserting (19) in Theorem 1 results in Corollary 1.■

When the system becomes unstable, the average response delay goes to infinity. Therefore, we define C^{stable} as the minimum stabilizer cache size for each SBS, where

$$C^{stable} = \min\{C \le F : \rho_n < 1 \ \ \forall n \in \boldsymbol{\mathcal{N}} \cup \{0\}\}\ . \tag{20}$$

Therefore, by using Propositions 1 and 2, we first obtain the value of C^{stable}, as the minimum cache size for having a stable system.

In the next step, using Theorem 1, we obtain C^* according to (21) which is the minimum acceptable cache size in order to have a tolerable delay.

$$C^* = \min\{C^{stable} \le C \le F : \overline{D} < D_{max}\}, \forall n \in \boldsymbol{\mathcal{N}}. \tag{21}$$

IV. Performance Evaluation through Trace Driven Experiment Results

In this section, the analytic expressions derived in this paper are validated through real trace-driven experiments on traffic of YouTube video requests. We have run a trace-driven experiment, using a real trace of video clips requests from a campus network measurement on YouTube traffic in 2008 [8], with a total 123.3k requests for 78.9k videos, arriving at 10 SBSs. We observe that the results achieved under synthetic traffic still hold when the cache is fed by real traffic taken from an operational network. We have compared the real trace results with the derived equations for IRM traffic. We have estimated the value of $\alpha = 0.55$ as the exponent parameter of the Zipf distribution for the popularity of the real trace requests. Figs. 2-5 study the performance of HCCNs from different aspects. In Fig. 2, the MBS and SBSs utilization factors are shown as functions of the cache size for different total requests throughputs, that is $\Lambda = \sum_{n=1}^{N} \lambda_{req}^{n}$, in case of the LFU scheme. The stable regions and values of C^{stable} are shown in this figure. We observe that by increasing the requests throughputs, the value of C^{stable} increases. In Fig. 3, we have shown the effect of the ratio of the file size to the downlink rate, i.e. $\frac{B}{r_0}$, on the values of C^{stable}. As shown in this figure, since Λ is less than one in the real trace, when the service rate is large enough, we can ensure the stability of the system for any cache size, that is, C^{stable} is zero in this case. However, by increasing the ratio of $\frac{B}{r_0}$, the values of C^{stable} increase rapidly. As shown in Fig. 3, the values of C^{stable} in case of LFU are explicitly less than LRU. Fig. 4 plots the average response delay as a function of the MBS and SBSs utilization factors. It shows that when ρ_0 becomes close to 1, the average response delay increases rapidly. This figure also shows that the effect of the MBS utilization factor on the average response delay is much more than that of the SBSs. Finally, Fig. 5 shows the average response delay as a function of the cache size. The minimum stable cache sizes and the effect of the requests throughputs are also illustrated in Fig. 5. As illustrated in this figure, the average response delay of LFU is less than LRU for different cache sizes. In addition, according to Fig. 5, increasing the cache size more than some values, does not significantly improve $\bar{D}$. Therefore, the network should be designed with a proper cache size according to the derivations and figures provided in this paper.

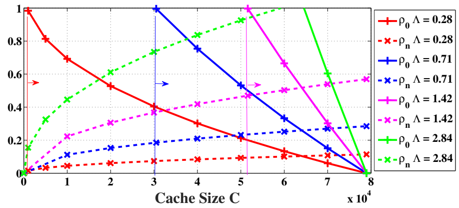

Fig. 2. MBS and SBSs utilization factors as a function of the cache size. Network parameters: $r_0 = r_n = 100\ Mbps$, B=400 Mb.

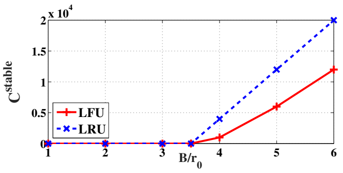

Fig. 3. Minimum stable cache size as a function of the ratio of the file size to the downlink bit rate, B/r_0.

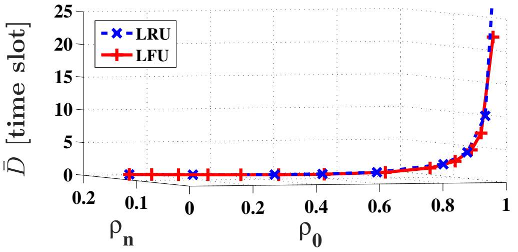

Fig. 4. Average response delay as a function of the MBS and SBSs utilization factors. Parameters: $r_0 = r_n = 100\ Mbps$, B=400 Mb.

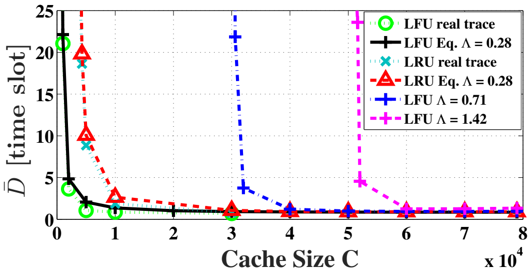

Fig. 5. Average response delay as a function of the cache size. Network parameters: $r_0 = r_n = 100\ Mbps$, B=400 Mb.

V. Conclusion and Future Works

In this paper, we have presented a delay and stability analysis in HCCNs. We have studied the effect of various network parameters on the delay performance and stability of such networks. We have also provided insight regarding the delay behavior in HCCNs and design of such networks in order to have tolerable delays. The extension of this work to the case of overlapping coverage regions will be studied in future works.

References

[1] X. Wang, M. Chen, T. Taleb, A. Ksentini, V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2. pp. 131-139, 2014.

[2] 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.*, vol. 51, no. 4, pp. 142–149, 2013.

[3] K. Poularakis, G. Iosifidis, L. Tassiulas, "Approximation Algorithms for Mobile Data Caching in Small Cell Networks", *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665-3677, 2014.

[4] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed Caching for Data Dissemination in the Downlink of Heterogeneous Networks", *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553-3568, 2015.

[5] F. Rezaei and B. H. Khalaj, "Stability, Rate and Delay Analysis of Single Bottleneck Caching Networks", *IEEE Trans. Commun.,* vol. 64, no.1, pp. 300-313, 2016.

[6] L. Kleinrock, *Queueing Systems: Vol. I*, New York: Wiley Interscience, 1975.

[7] V. Martina, M. Garetto, and E. Leonardi, "A unified approach to the performance analysis of caching systems," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM'14),* 2014, pp. 2040–2048.

[8] M. Zink, K. Suh, Y.Gu and J.Kurose, "Characteristics of YouTube network traffic at a campus network -Measurements, models, and implications", *Int. J. Comput. Telecommun. Netw.,* vol. 53 no. 4, pp. 501–514, 2009.