Performance Analysis of Heterogeneous Cellular Caching Networks with Overlapping Small Cells

Fatemeh Rezaei, Babak H. Khalaj, Ming Xiao, and Mikael Skoglund

***Abstract*—Caching at network edges has attracted more and more research interests recently for the purpose of alleviating the network traffic pressure especially in backhaul links and improving user experience. We study Heterogeneous Cellular Caching Networks (HCCNs) consisting of macro cells in which N small cell base stations (SBSs) equipped with cache memory operate in conjunction with the macro cell base station (MBS). We provide closed-form expressions of the MBS and SBSs utilization factors and average user-experienced-delay in HCCNs with overlapping coverage regions, considering general traffic models for the request arrivals based on the Independent Reference Model (IRM) and renewal traffic models. Moreover, we propose a novel caching scheme in HCCNs, namely Cooperative Most Popular Caching (CMPC), which outperforms the existing schemes in terms of delay. Subsequently, we present the bandwidth assignment problem aiming to minimize the average user-experienced-delay under stability and cache size constraints in HCCNs with overlapping coverage regions and stochastic request arrivals. Finally, the analytic results are validated through numerical results and real trace-driven experiments.**

***Index Terms*—heterogeneous cellular networks, caching, queuing analysis, latency, overlapping small cells.**

I. Introduction

DATA traffic generated by wireless and mobile devices has dramatically increased in recent years. To meet the increasing demand for higher data rates and lower delay, various approaches have been proposed such as heterogeneous networks (HetNets) and network caching [1]. Network caching has attracted more and more research interests in recent years to alleviate the communication traffic pressure in networks and improve user experience. It was shown numerically up-to 66 % reduction in network traffic by using caching in 3G [2] and 4G [3] networks. Potential techniques for caching in 5G mobile networks, including evolved packet core network caching and radio access network caching, were also studied in [4], [5].

A. Related work

The problem of caching in Heterogeneous Cellular Networks (HCNs) has been recently addressed from different perspectives. For instance, the authors in [6] studied the effect of retransmissions on the optimal cache placement policy and determined the optimal caching probability of the files that maximize the hit probability. The virtual resource allocation strategy as a joint optimization problem in the information-centric heterogeneous networks was proposed in [7]. The authors in [8] formalized the delay minimization problem by assuming that users can directly obtain files from the macro cell base station (MBS) with the maximum delay and showed that the problem is NP-hard. Distributed caching optimization algorithms via belief propagation (BP) for minimizing the downloading latency in HCNs were proposed in [9].

It should be noted that the related works in [6]-[14] did not consider stochastic request arrivals. However, in practical networks, the stochastic arrival of user requests should be considered to analyze the impact on the performance. For this purpose, we introduced a framework based on queuing theory in single bottleneck caching networks considering the stochastic arrival of user requests in our previous works [15], [16]. We provided the analysis of the stability, throughput, load on the bottleneck link, and delay for various caching schemes in such networks [16]. We proved that the coded caching schemes in the literature lead to unstable systems in the case of stochastic arrivals of user requests. In [17], we studied Heterogeneous Cellular Caching Networks (HCCNs) with stochastic request arrivals and addressed the problem of finding the minimum cache size for the SBSs to achieve a tolerable average delay. The main drawback of our previous work in [17] is the assumption that there is no overlap between the SBSs coverage regions where each user accesses only one SBS. Afterward, [18]-[22] studied cache enabled HCNs by considering stochastic request arrivals.

The authors in [18] analyzed the cache-based content delivery in a heterogeneous network, where each of the relays has the same size caching storage and proactively caches the same copy of the most popular content when the network is off-peak. However, the detailed impacts of the limited caching space and content popularity were not considered. Reference [19] studied caching control and the bandwidth allocation problem aims at minimizing the request miss ratio in a heterogeneous small-cell caching system consisting of one MBS and nonoverlapping SBSs. They assumed Poisson arrivals, deterministic service time, and zero-length buffer. Considering the binary caching decisions of all BSs, they formulated the caching problem that aims to minimize the request miss rate. The authors in [20] formulated a stochastic content multicast scheduling problem to jointly minimize the average network delay and power costs under a multiple access constraint. Reference [21] formulated the cooperative content caching as an integer

F. Rezaei is with the Department of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran. B. H. Khalaj is with the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran. M. Xiao and M. Skoglund are with the Department of Information Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. e-mails:(frezaei@kntu.ac.ir, khalaj@sharif.edu, mingx@kth.se, skoglund@kth.se.)

linear programming problem aiming to minimize the average downloading latency in HCNs. In [20], [21], it was assumed that requests arrive independently following Poisson processes.

In general, related works studied mechanisms in which files are proactively cached during off-peak demands. However, in practical networks, online caching schemes based on the stochastic arrival of the user requests should be considered. Related works studying delay minimization in HCCNs did not consider online caching schemes, hit probabilities, and stability constraints. Moreover, they only considered binary caching decision variables and provided combinatorial optimization problems with integer variables that were NP-hard and led to approximation algorithms.

B. Contributions

In this paper, we study HCCNs consisting of macro cells in which N SBSs equipped with cache memory operate in conjunction with the MBS. We consider a general and practical structure for the network architecture with overlapping coverage regions where users access an arbitrary number of the SBSs. We propose a practical network framework considering the stochastic arrival of the user requests, for both online and proactive caching schemes. The main contributions of this paper are summarized as follows:

- We propose a new framework for analyzing the role of caching in general HCNs with overlapping coverage regions from a queuing theory perspective, by considering the stochastic request arrivals based on the Independent Reference Model (IRM) and renewal traffic models.
- We provide closed-form expressions of the MBS and SBSs utilization factors and the average experienced delay and illustrate the relation between prior network metrics such as the hit probability, link load, and latency in HCCNs with overlapping coverage regions and stochastic request arrivals.
- We formulate the bandwidth assignment problem aiming to minimize the average user-experienced-delay under the stability and cache size constraints in the studied network.
- We propose a novel caching scheme in the practical model of HCCNs with overlapping coverage regions, namely Cooperative Most Popular Caching (CMPC), which outperforms the existing schemes in terms of latency.

The rest of the paper is organized as follows. Section II describes the system model. In Section III, the performance analysis of HCCNs is presented. A novel cooperative low latency caching scheme and the bandwidth assignment problem are proposed in Section IV. In Section V, the performance evaluation through numerical results is presented. Finally, Section VI concludes the paper.

II. System Model

We study HCNs wherein each macro cell, multiple SBSs operate in conjunction with the MBS. We consider overlapping coverage regions where users access an arbitrary number of the SBSs as illustrated in Fig. 1. For notational convenience, a single macro cell is considered, which is easily extended to multiple macro cells. The set of mobile users (MUs) submitting their content requests to the mobile network operator (MNO) is denoted by $\boldsymbol{M} = \{m_1, m_2, \cdots, m_U\}$. Moreover, $\boldsymbol{S} = \{s_1, s_2, \cdots, s_N\}$ denotes the set of the SBSs where $s_n, n \in \boldsymbol{N} = \{1, 2, \cdots, N\}$ represents the nth SBS.

We assume that the requests are drawn from a specific same-size file library $\mathbb{F} = \{f_i\}, i \in \boldsymbol{F} = \{1, \cdots, F\}$ of size B bits, for notational convenience, similar to [22]-[24]. This assumption can be easily removed by dividing the content items into multiple smaller files of the same length [21], [25]. We consider a general model for the library file popularity, where the probability of requesting file $f_i, i \in \boldsymbol{F}$, is denoted by p_i. The cache content of SBS s_n, denoted by z_n, is a subset of the library $\mathbb{F}$. SBS s_n is capable of storing C_n complete files (i.e., $C_n B$ bits). For notational convenience, we also define s_0 denoting a virtual SBS with a zero-size cache, i.e., $C_0 = 0$, for the users that only access the MBS without any access to any SBSs.

The hit probability for file f_i at s_n, i.e., $prob\,(f_i \in z_n)$, is denoted by $p_{hit}(i, n)$. By definition, the hit probability is the limit of the hit ratio when the number of requests goes to infinity. The performance of different caching schemes can be modeled by their hit probabilities [16], [26]. In [16], we proposed hit probabilities for LRU, q-LRU, LFU, and RAND [27] caching schemes in terms of the network parameters.

Different locations of the MUs in the macro cell, path loss, and channel fading cause random channel conditions and downlink rates. 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].

To properly model and analyze the overlapping coverage regions of the SBSs, we group the MUs and SBSs such that the MUs in each group of the users only access the SBSs in a corresponding group of the SBSs. It should be noted that based on the network topology and the overlapping coverage ranges of the SBSs, the macro cell is divided into $G+1$ zones, such that the MUs in each zone have only access to a specific subset of the SBSs. In more detail, the set of the MUs, $\boldsymbol{M}$, is partitioned into $G+1$ subsets, where $M_g \subseteq \boldsymbol{M}, g \in \boldsymbol{G} = \{0, 1, 2, \cdots, G\}$ represents the gth subset. The subset of the accessible SBSs in the gth zone is also denoted by $\overline{\mathrm{S}}_g \subseteq \boldsymbol{S}, g \in \boldsymbol{G}$. In other words, the grouping is set such that the MUs in M_g are in the coverage area of the SBSs in $\overline{\mathrm{S}}_g$ and only have access to them. For notational convenience, the subset M_0 represents the users that only access the MBS without any access to any SBSs, in correspondence with the SBS subset $\overline{\mathrm{S}}_0 = \{s_0\}$. For example, consider a network with $N = 7$ SBSs and 9 groups of mobile users ($G = 8$), as in Fig. 1. The solid curves represent the user groups, i.e., $M_g, g \in \{0, 1, 2, \cdots, 8\}$, and the dashed circles show the SBS coverage ranges. In this figure, the users in M_1 are only in the coverage of SBS s_1, hence $\overline{\mathrm{S}}_1 = \{s_1\}$. The users in M_2 are in the coverage of SBSs s_2 and s_3, hence $\overline{\mathrm{S}}_2 = \{s_2, s_3\}$, and so on.

In this paper, we consider stochastic request arrivals and present the equations for the general traffic of requests, where the average request arrival rate of the users in M_g is denoted

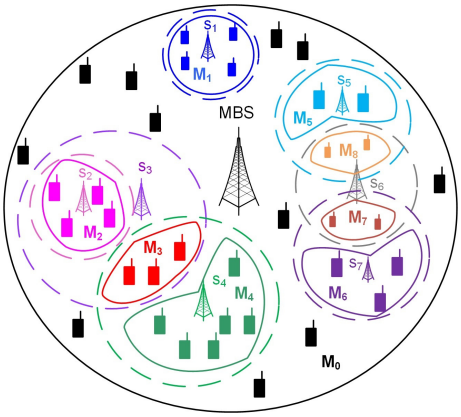


Fig. 1. Macro cell model, $N = 7, G = 8$. The solid curves represent the user groups, $M_g, g \in \{0, 1, 2, \cdots, 8\}$. The dashed circles show the SBS coverage ranges. The SBS groups for this topology: $\overline{S}_1 = \{s_1\}, \overline{S}_2 = \{s_2, s_3\}, \overline{S}_3 = \{s_3, s_4\}, \overline{S}_4 = \{s_4\}, \overline{S}_5 = \{s_5\}, \overline{S}_6 = \{s_7\}, \overline{S}_7 = \{s_6, s_7\}, \overline{S}_8 = \{s_5, s_6\}$.

TABLE I
KEY NOTATIONS FOR THE SYSTEM MODEL AND ANALYSIS

Notation	Semantics [unit]
N	Number of SBSs
F	Size of the file library that users can request files from
p_i	Probability of requesting file f_i from the library
C_n	Cache size of SBS s_n [number of whole files]
B	File size [bits]
τ	Length of the transmission time slot [sec]
r_n	Average downlink rates [bps]
$p_{hit}(i,n)$	Hit probability for file f_i at s_n
p_L	Probability of serving a requested file locally by SBSs in the case of hitting
λ^g_{req}	Average request arrival rate in the user group M_g [files/second]
λ_n	Average arrival rate at the queues [files/time slot]
μ_n	Average service rate at the queues [files/time slot]
ρ_n	The MBS and SBSs utilization factors

by λ^g_{req} [files per second].

Definition 1: We define $p_{req}(i,g)$ as the probability that file f_i is requested by the users in M_g within a transmission time slot of length τ. We also define the random variable $X_{req}(i,g)$ as the event of requesting file f_i in the user group M_g within a time slot.

Definition 2: We define p_L as the probability that a requested file is locally served by an SBS in the case of hitting.

According to the model, we consider a serving probability, denoted by $p_{serv}(i,n)$ representing the probability that a requested content f_i is served by s_n. Based on definitions, the probability that the requested content f_i is served by the SBS s_n is given by

$$p_{serv}(i,n) = p_L p_{hit}(i,n). \tag{1}$$

In such a network, the MUs submit their requests to the MNO. If the requested content for users in M_g is present (hits) in a cache of the SBSs in $\overline{S}_g$, the MNO sends the request to the SBSs with probability p_L and the request is served locally.[1] Otherwise, the content is sent to the user directly from the MBS. Similar to [25], we consider the case that each request is entirely served by one base station. That is, a user that requests a file will not receive different parts of it from different base stations. This assumption is important, since associating a user request with multiple base stations incurs the extra effort to synchronize the communication and it may lead to network instability as studied in [16]. The main network parameters and notations are given in Table I.

III. PERFORMANCE ANALYSIS

In this section, we will propose a performance analysis approach for HCCNs with overlapping coverage regions from a queuing theory perspective. Since the channel from each SBS to the MUs is shared among the users, there is competition among users to receive their requested files via that downlink path. On the other hand, the requests that are not served locally by the SBSs enter the MBS, where the channel from the MBS to the users is also shared. Therefore, we model the function of the MBS and SBSs by controlled first-in-first-out (FIFO) queues, $Q_n, n \in \boldsymbol{N} \cup \{0\}$, where control units ensure that when multiple users request the same file concurrently (requests overlap within a time slot), the MBS and each SBS only store the file in a single location of their individual queues. The average request arrival rates and service rates at the MBS and SBSs transmission queues are denoted by λ_n and μ_n for $n \in \boldsymbol{N} \cup \{0\}$, respectively. Since the downlink rates for serving the requested contents are varying over time, due to the random channel conditions, the service time is random and is modeled by a general (arbitrary) distribution, with mean and standard deviation, $\frac{1}{\mu_n}$ and σ_n, respectively. Therefore, we consider general G/G/1 queue models, where the inter-arrival and service times have arbitrary distributions. In the case of independent reference model (IRM) traffic [26], [28], the MBS and SBSs functionalities are modeled by M/G/1 queues [29]. By means of the proposed queue models, we derive the MBS and SBSs utilization factors, i.e., $\rho_n = \frac{\lambda_n}{\mu_n}, n \in \boldsymbol{N} \cup \{0\}$. The key advantage of such metrics is that they simultaneously take into account the cache hit probabilities, load on the links, and requests arrival rates. Moreover, they provide valuable insight into the stability characteristic and delay behavior of HCCNs.

A. The utilization factors of the SBSs

In this section, we propose the wireless link utilization factor of the cache-enabled SBSs for the network model presented in Section II. It should be noted that in the case of overlapping coverage regions of the SBSs, the analysis is much more complicated, and to derive the link utilization factor of each SBS, we have to take into account the file requests from all of the users' subsets who are in the coverage area of this SBS. In order to precisely derive the link utilization factor of each SBS, we define $p_{sbs.req}(i,n)$ as the probability that file f_i is requested within a time

[1] It should be noted that even in cases that a requested file is available locally in an SBS cache, we might incur less delay in certain scenarios by getting the file directly from the MBS. It might result from large values of the SBS queue utilization factor that will be studied further in the next sections. So a value of p_L less than one might be adapted.

slot in at least one of the user groups in the coverage range of the SBS s_n. Based on the definitions, we have $p_{sbs.req}(i,n) = P\big(\bigcup_{g:s_n\in\overline{\mathrm{S}}_g} X_{req}(i,g)\big)$.

Proposition 1: The utilization factor of the SBS s_n wireless link in HCCNs with overlapping coverage regions is obtained from

$$\rho_n = \frac{Bp_L}{r_n\tau}\sum_{i=1}^{F} p_{hit}(i,n)\, p_{sbs.req}(i,n),\ n\in\boldsymbol{N}, \tag{2}$$

and in case that request arrivals in different user groups are independent, we have

$$\rho_n = \frac{Bp_L}{r_n\tau}\sum_{i=1}^{F} p_{hit}(i,n)\left(1-\prod_{g:s_n\in\overline{\mathrm{S}}_g}\big(1-p_{req}(i,g)\big)\right),\ n\in\boldsymbol{N}. \tag{3}$$

Proof: See Appendix A.

In Proposition 1, we have derived the wireless link utilization factor of the SBSs in the general form and for any arbitrary caching policy and requests traffic. The key advantage of this proposition is that it simultaneously takes into account the cache hit probabilities, average downlink rates, and requests traffic in the general form. In the following lemma, we present the probability of requesting each file in each user subset for two types of important and practical traffic models which can be applied in Proposition 1.

Definition 3: We define the random variable $N_{\tau,i,g}$, as the number of requests for file f_i coming from the user group M_g within a time slot of length τ.

Lemma 1: For the renewal traffic model [30], $p_{req}(i,g)$ is obtained from

$$p_{req}(i,g) = 1 - G_{i,g}(\tau,0)\,, \tag{4}$$

where $G_{i,g}(\tau,\xi)$ is the probability generating function of $N_{\tau,i,g}$. For the IRM traffic model, $p_{req}(i,g)$ is obtained from

$$p_{req}(i,g) = 1 - e^{-\lambda^g_{req}p_i\tau}. \tag{5}$$

Proof: See Appendix B.

Using Lemma 1, we present the link utilization factor of the cache-enabled SBSs in case that the traffic of the users' requests in the coverage area of the SBSs is modeled by the IRM model.

Corollary 1: The utilization factor of the SBS s_n wireless link in HCCNs for the IRM traffic is given by

$$\rho_n = \frac{Bp_L}{r_n\tau}\sum_{i=1}^{F} p_{hit}(i,n)\left(1-e^{-\sum_{g:s_n\in\overline{\mathrm{S}}_g}\lambda^g_{req}p_i\tau}\right),\ n\in\boldsymbol{N}. \tag{6}$$

Proof: It is simply driven according to Lemma 1 and Proposition 1.

Therefore, by substituting the cache hit probabilities, average requests arrival rates, and average downlink rates in Proposition 1 and Corollary 1, we obtain the link utilization factor of the SBSs. So, we can determine the system stability according to the following definition.

Definition 4: A packet queue is stable if the average arrival rate is smaller than the average service rate [31].

By definition, to have a stable system, we require $\lambda_n < \mu_n$, that is $\rho_n < 1$, for $\forall n \in \boldsymbol{N}\cup\{0\}$.

B. The utilization factor of the MBS

According to the HCCN operation framework discussed earlier, the requested contents of the users in M_g that are not served in the caches of the SBSs in $\overline{\mathrm{S}}_g$, are served by the MBS. In more detail, if at least one user in one user group requests file f_i and none of the SBSs serves it, this file enters the MBS queue. Because of the complexities in the case of overlapping coverage regions of the SBSs and the fact that each user may obtain the requested file from different SBSs, it is important to precisely model and derive the MBS link utilization factor. The following definitions and lemmas let us precisely present the MBS link utilization factor in Proposition 2.

Definition 5: We define $p_{joint.req}(i,J)$ as the probability that file f_i is requested within a time slot in all of the user groups $M_g : g\in J$, for any subset $J\subseteq\boldsymbol{G}$.

Lemma 2: $p_{joint.req}(i,J)$ is obtained from

$$p_{joint.req}(i,J) = P\left(\bigcap_{g\in J} X_{req}(i,g)\right). \tag{7}$$

Proof: It directly results from the definitions.

Definition 6: For any subset $J\subseteq\boldsymbol{G}$, we define $p_{unserv}(i,J)$ as the probability that a request for file f_i is not served in the caches of the SBS groups $\overline{\mathrm{S}}_g : g\in J$. We also define the random variable $X_{unserv}(i,n)$ as the event of not serving the request for file f_i in the cache of SBS s_n.

Lemma 3: $p_{unserv}(i,J)$ is obtained from

$$p_{unserv}(i,J) = P\left(\bigcap_{n:s_n\in\left(\bigcup_{g\in J}\overline{\mathrm{S}}_g\right)} X_{unserv}(i,n)\right). \tag{8}$$

Proof: It directly results from the definitions.

Proposition 2: The utilization factor of the MBS link in HCCNs with overlapping coverage regions is derived as

$$\rho_0 = \frac{B}{r_0\tau}\sum_{i=1}^{F}\sum_{|J|=1}^{|\boldsymbol{G}|}(-1)^{|J|+1}\sum_{J\subseteq\boldsymbol{G}} p_{joint.req}(i,J)\, p_{unserv}(i,J), \tag{9}$$

where $p_{joint.req}(i,J)$ and $p_{unserv}(i,J)$ are given by Lemmas 2 and 3, respectively.

Proof: See Appendix C.

In Proposition 2, we have proposed ρ_0 in general. In order to make the presentation simpler, we propose ρ_0 in a special case in the following corollary.

Corollary 2: If *a) The request arrivals in different user groups are independent*, and *b) The requests are independently served at different SBSs, or the requests are almost surely*

served or not served at the SBSs, the utilization factor of the MBS wireless link in HCCNs is derived as

$$\rho_0 = \frac{B}{r_0\tau}\sum_{i=1}^{F}\sum_{|J|=1}^{|\boldsymbol{G}|}(-1)^{|J|+1}\sum_{J\subseteq\boldsymbol{G}}\prod_{g\in J}p_{req}(i,g)\prod_{n:s_n\in\left(\cup_{g\in J}\overline{\mathrm{S}}_g\right)}\big(1-p_L p_{hit}(i,n)\big). \quad (10)$$

Proof: See Appendix D.

It should be noted that the aforementioned assumptions in Corollary 2 can appear in practical networks because different user groups are distinct elements in the network and we can assume that the request arrivals at different groups are independent. In addition, since the network operator designs the network, we can apply scenarios in which the requests are independently served at different SBSs, or for example, the files are proactively cached at the SBSs.

So far, we have presented the general form and also practical special cases of the SBSs and MBS link utilization factors based on the cache hit probabilities, request probabilities, and system parameters in the studied network. We will investigate and compare the behavior of these metrics in practical scenarios in Section V.

C. Delay analysis

In this section, we present the average user-experienced-delay, $\overline{D}$, defined as the average delay experienced by *any given user* in HCCNs for obtaining the requested files, either delivered locally from the SBSs or sent directly from the MBS. To obtain the average experienced delay for any given user in user subset M_g, we should consider the case that the requested files are delivered locally from the accessible SBSs, as well as the case that the requested files are directly delivered from the MBS.

We define $p_{serv}(g)$ as the probability that any requested file from any given user in M_g is served by the SBSs in $\overline{\mathrm{S}}_g$. Considering the library file popularity $p_i, i \in \boldsymbol{F}$, and averaging over all of the requested files in the library, we have

$$p_{serv}(g) = 1 - \sum_{i=1}^{F} p_i \prod_{n:s_n\in\overline{\mathrm{S}}_g}\big(1-p_L\, p_{hit}(i,n)\big), \quad (11)$$

where the right-hand side results from (1). Therefore, the average experienced delay for any given user in user subset M_g is obtained from

$$\overline{D}_g = p_{serv}(g)\,\overline{dl}_g + \big(1-p_{serv}(g)\big)dl_0, \quad (12)$$

where $\overline{dl}_g = \frac{1}{|\overline{\mathrm{S}}_g|}\sum_{n:s_n\in\overline{\mathrm{S}}_g} dl_n, g \in \boldsymbol{G}$, is the average delay from the SBSs in $\overline{\mathrm{S}}_g$, and $dl_n, n \in \boldsymbol{N}\cup\{0\}$, denotes the average delay from the SBSs and MBS. Finally, by averaging the experienced delay in all of the user subsets, the average user-experienced-delay in the network is given by

$$\overline{D} = \frac{1}{|\boldsymbol{G}|}\sum_{g\in\boldsymbol{G}}\overline{D}_g. \quad (13)$$

In the following proposition, we present the average user-experienced-delay in HCCNs with overlapping coverage regions, as a function of the cache hit probabilities, link utilization factors, and network parameters.

Proposition 3: In the case of the IRM traffic of the user requests, $\overline{D}$ is obtained from

$$\overline{D} = \frac{1}{|\boldsymbol{G}|}\sum_{g\in\boldsymbol{G}}\Bigg(\Bigg(1-\sum_{i=1}^{F}p_i\prod_{n:s_n\in\overline{\mathrm{S}}_g}\big(1-p_L\,p_{hit}(i,n)\big)\Bigg)\times\Bigg(\frac{1}{|\overline{\mathrm{S}}_g|}\sum_{n:s_n\in\overline{\mathrm{S}}_g}\frac{B}{r_n\tau}\Big(1+\frac{\rho_n\big(1+\upsilon_n^2\big)}{2(1-\rho_n)}\Big)\Bigg)+\Bigg(\sum_{i=1}^{F}p_i\prod_{n:s_n\in\overline{\mathrm{S}}_g}\big(1-p_L\,p_{hit}(i,n)\big)\Bigg)\frac{B}{r_0\tau}\tau\Big(1+\frac{\rho_0\big(1+\upsilon_0^2\big)}{2(1-\rho_0)}\Big)\Bigg), \quad (14)$$

where $\upsilon_n = \mu_n\sigma_n$ is the coefficient of the variation of the service time.

Proof: See Appendix E.

Given the derived formulas for the utilization factors and delay in HCCNs with overlapping coverage regions, in the next section, we will study performance improvement in such networks.

IV. Performance Improvement and Insights

According to the derived formulas in the previous section, we will provide some insights on the performance of HCCNs with overlapping coverage areas of the SBSs. It should be noted that in each macro cell, there are one MBS and N number of SBSs. In addition, all of the mobile users are in the coverage of the MBS, but only a portion of the MUs, who are in the user groups $M_g, g : s_n \in \overline{\mathrm{S}}_g$, are in the coverage area of the SBS s_n. Therefore, the requests coming to each SBS are a fraction of the whole requests in the macro cell. Moreover, the MBS contains all of the files in the library, but each SBS only caches a small fraction of the library files. Consequently, in practical networks, λ_n is a small fraction of λ_0. According to definition, $\rho_n = \frac{\lambda_n}{\mu_n}, n \in \boldsymbol{N}\cup\{0\}$, ρ_n is proportional to λ_n, and ρ_0 is proportional to λ_0. This fact is also illustrated in Propositions 1 and 2. In Proposition 1, it is illustrated that only the user requests in $M_g, g : s_n \in \overline{\mathrm{S}}_g$ are considered in ρ_n. In Proposition 2, it is shown that all of the user requests in the macro cell, i.e., $M_g, g \in \boldsymbol{G}$, are considered in ρ_0.

On the other hand, in practical upcoming networks, such as non-standalone development of 5G mobile networks, the average downlink rate of 5G SBSs is much greater than the average downlink rate of 4G MBSs. Therefore, the SBS service time, $\frac{1}{\mu_n}, n \in \boldsymbol{N}$ is smaller and negligible compared to the MBS service time. Therefore, in practical networks, it is reasonable to consider the case of $\rho_n \ll \rho_0$. According to Definition 4, to have a stable network, the utilization factors must be smaller than one, i.e., $\rho_n \ll \rho_0 < 1$. Therefore, the stability of the MBS is the bottleneck in such networks.

Algorithm 1: Cooperative Most Popular Caching scheme (CMPC)

$\mathbf{\Psi}_g$:The pool of the gth group containing $\sum_{n:s_n\in\overline{\mathrm{S}}_g} C_n$most popular files
while $\exists s_n \in \big(\underset{g\in\boldsymbol{G}}{\cup}\overline{\mathrm{S}}_g\big)$, **do**
  $i = \underset{g\in\boldsymbol{G}:|\overline{\mathrm{S}}_g|\neq\emptyset}{min}|\overline{\mathrm{S}}_g|$
  while $\exists s_n \in \big(\underset{g:|\overline{\mathrm{S}}_g|=i}{\cup}\overline{\mathrm{S}}_g\big)$, **do**
    Select s_{n^*} **for cache placement:**
      choose one SBS belonged to the largest number of groups $\overline{\mathrm{S}}_g$
    Select the pool of the files $\mathbf{\Psi}_{g^*}$:
      $g^* = \underset{g:s_{n^*}\in\overline{\mathrm{S}}_g}{argmax}|\mathbf{\Psi}_g|$
    Cache placement for z_{n^*}:
      $z_{n^*} \leftarrow$ assign C_{n^*} of the most popular files of the $\mathbf{\Psi}_{g^*}$
    Update SBSs Groups:
      for $\forall g : s_{n^*} \in \overline{\mathrm{S}}_g$
        $\mathbf{\Psi}_g = \mathbf{\Psi}_g \setminus z_{n^*}$
        $\overline{\mathrm{S}}_g = \overline{\mathrm{S}}_g \setminus \{s_{n^*}\}$
      end for
  end while
end while

A. Proposing a cooperative low latency caching scheme in HCCNs

Now let's have a look at the delay performance. According to (12), (13), we can write

$$\overline{D} = \frac{1}{|\boldsymbol{G}|}\sum_{g\in\boldsymbol{G}}\Big(p_{serv}(g)\,\overline{dl}_g + \big(1-p_{serv}(g)\big)dl_0\Big). \tag{15}$$

So, when the SBS service time, $\frac{1}{\mu_n}, n\in\boldsymbol{N}$, approaches zero, consequently, $\overline{dl}_g = \frac{1}{|\overline{\mathrm{S}}_g|}\sum_{n:s_n\in\overline{\mathrm{S}}_g}\frac{1}{\mu_n}\big(1+\frac{\rho_n(1+\upsilon_n^2)}{2(1-\rho_n)}\big)$ approaches zero. Therefore, to minimize the average user-experienced-delay in HCCNs under the cache size constraints, for the IRM traffic, and in case that the SBS service time, $\frac{1}{\mu_n}, n\in\boldsymbol{N}$, approaches zero, we should minimize

$$\frac{1}{|\boldsymbol{G}|}\sum_{g\in\boldsymbol{G}}\big(1-p_{serv}(g)\big)dl_0 = \frac{1}{|\boldsymbol{G}|}\sum_{g\in\boldsymbol{G}}\big(1-p_{serv}(g)\big)\frac{1}{\mu_0}\big(1+\frac{\rho_0(1+\upsilon_0^2)}{2(1-\rho_0)}\big). \tag{16}$$

Maximizing $p_{serv}(g)$ results in minimizing $\rho_0, dl_0 = \frac{1}{\mu_0}\big(1+\frac{\rho_0(1+\upsilon_0^2)}{2(1-\rho_0)}\big)$, and subsequently (16). By caching the popular files in the SBSs, $p_{serv}(g)$ increases. The performance bound in the literature is obtained based on the most popularity caching strategy as the optimal caching scheme for each SBS [18], [19]. However, in the Least Frequently Used (LFU) caching scheme [27], each cache contains the most popular files in the library and therefore, there is redundancy between the cache contents of different SBSs in a group. In addition, $\underset{n:s_n\in\overline{\mathrm{S}}_g}{max}\,C_n$ most popular files are cached in each group $\overline{\mathrm{S}}_g$ in LFU.

It should be noted that in the case of overlapping coverage regions, one SBS may be in more than one SBS group. Therefore, we need an efficient cooperative algorithm for performance improvement. To achieve this goal, we propose a caching algorithm in HCCNs, namely Cooperative Most Popular Caching (CMPC) scheme. In Algorithm 1, all the caches in $\overline{\mathrm{S}}_g$ are considered together as a single distributed cache with the maximum size of $\sum_{n:s_n\in\overline{\mathrm{S}}_g} C_n$. More specifically, we consider a pool of $\sum_{n:s_n\in\overline{\mathrm{S}}_g} C_n$ most popular files in each group $\overline{\mathrm{S}}_g$. For cache placement, the smallest sizes of the SBSs groups are selected first. Among the SBSs in groups of size i, one SBS that is belonged to the largest number of groups $\overline{\mathrm{S}}_g$ is chosen. After selecting SBS s_{n^*} for cache placement, the largest pool of the groups containing s_{n^*}, i.e., $\mathbf{\Psi}_{g^*}$ is selected. Then, C_{n^*} of the most popular files of $\mathbf{\Psi}_{g^*}$ are cached in s_{n^*}. Subsequently, the cached files are removed from the pool. This procedure continues until caching the whole SBSs. For instance, in the network illustrated in Fig. 1, by running the proposed algorithm, $s_4, s_5, s_7, s_1, and\ s_3$ cache the most popular files under their cache size constraints. The files from $min(C_5, C_7)$ till $min(C_5, C_7)+C_6$ cache in s_6. Finally, the files from C_3 till $C_3 + C_2$ cache in s_2. Depending on the network topology and SBSs groups, the proposed scheme improves the performance in terms of latency. The performance of the CMPC scheme in the worst case is the same as the LFU scheme. The complexity of the proposed algorithm is $O(n)$, where the input size is the number of user requests. The proposed CMPC scheme is also a suboptimal solution for delay minimization in practical networks, even without the assumption of small average service times of the SBSs. This issue is verified in Section V by simulations and real trace-driven experiments.

B. Bandwidth assignment for delay minimization in HCCNs

From the design perspective, besides selecting the proper cache placement strategy in HCCNs, the downlink rate of the BSs will have a key role in the network performance. Similar to other works in HCNs [25], [32], we have considered orthogonal channels from the SBSs to the MUs with the average downlink rates denoted by r_n [bps]. Therefore, one important question coming up next is that if we have a bandwidth constraint BW [bps], how the bandwidth assignment for each SBS should be considered in each group. Based on the derivations in Section III, we address the problem of bandwidth assignment for different SBSs, in Proposition 4:

Proposition 4: The optimum bandwidth assignment for different SBSs, in order to minimize the average user-experienced-delay, is the solution of the following convex optimization problem:

$$minimize\ \overline{D} = \frac{1}{|\boldsymbol{G}|}\sum_{g\in\boldsymbol{G}}\frac{p_{serv}(g)}{|\overline{\mathrm{S}}_g|}\sum_{n:s_n\in\overline{\mathrm{S}}_g}\frac{B}{\tau r_n}\Big(1+\frac{\frac{B\lambda_n\left(1+\upsilon_n^2\right)}{\tau r_n}}{2\big(1-\frac{B\lambda_n}{\tau r_n}\big)}\Big) + \big(1-p_{serv}(g)\big)\frac{B}{\tau r_0}\big(1+\frac{\rho_0\left(1+\upsilon_0^2\right)}{2(1-\rho_0)}\big), \tag{17}$$

$$s.t.$$

$$\sum_{n:s_n\in\overline{\mathrm{S}}_g} r_n - BW = 0, \quad \forall g\in\boldsymbol{G}, \tag{18}$$

$$\frac{B\lambda_n}{\tau} - r_n < 0, \quad \forall n\in\boldsymbol{N}, \tag{19}$$

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3132648

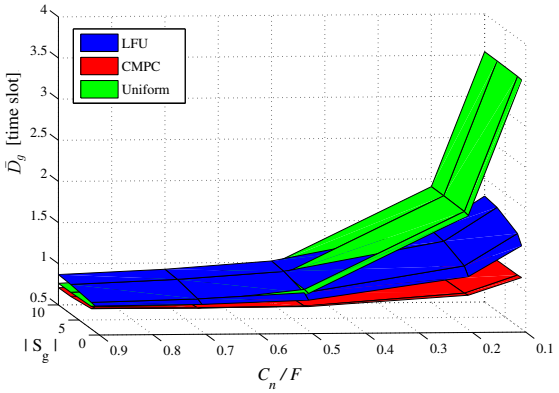


Fig. 2. The average delay of different caching schemes in HCCNs as functions of the SBSs group size and cache size. Network parameters: $F = 78.9k$, $N = 20$, $G = 4$, $B = 70$ Mb, $r_0 = r_n = 100$ Mbps.

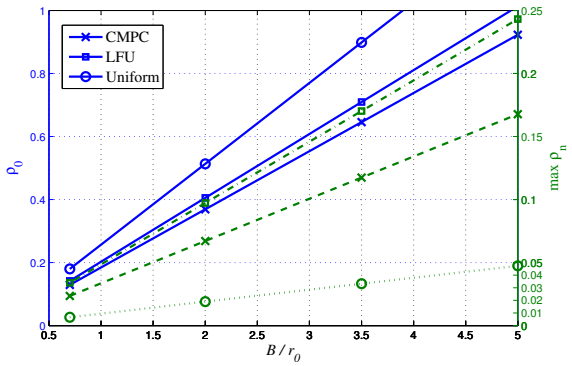


Fig. 4. The MBS and SBSs utilization factors of different caching schemes as functions of the ratio of the average file size to the downlink rate. Network parameters: $F = 78.9k$, $N = 4$, $G = 3$, $r_n = r_0$, $C_n = 0.1F$.

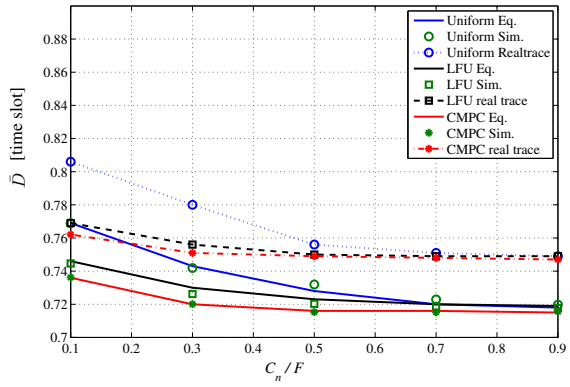


Fig. 3. The simulation, equations and real trace results of the average experienced delay of different schemes for variable cache sizes. Network parameters: $F = 78.9k$, $N = 4$, $G = 3$, $B = 70$ Mb, $r_0 = r_n = 100$ Mbps.

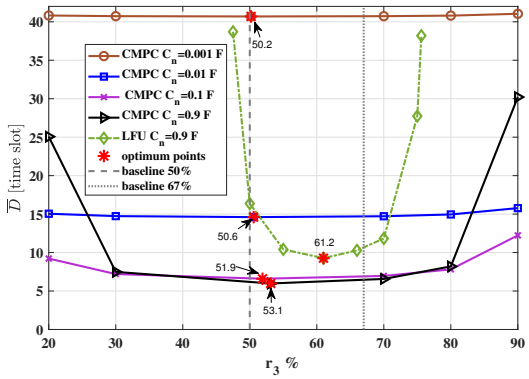


Fig. 5. The bandwidth assignment for variable cache sizes. Network parameters: $F = 78.9k$, $N = 4$, $G = 3$, $B = 350$ Mb, r_0 =BW=100 Mbps.

where $\lambda_n = \sum_{i=1}^{F} p_{serv}(i,n)\left(1 - e^{-\sum_{g:s_n \in \bar{S}_g} \lambda_{req}^{g} p_i \tau}\right)$, and (19) is the stability constraint.

Proof: See Appendix F.

Since the problem in Proposition 4 is a convex optimization problem with linear constraints, it is solved in polynomial time [33]. Subsequently, the optimum bandwidth assignment for different schemes is obtained as presented in Section V.

V. Performance Evaluation through Numerical Results

In this section, the analytic expressions derived in this paper are validated through simulation results and real trace-driven experiments on the traffic of YouTube requests. We have run a trace-driven experiment, using a real trace of requests from a campus network measurement on YouTube traffic in 2008 [34], with a total $123.3k$ requests for $F = 78.9k$ files. Figs. 2-5 study the performance of HCCNs from different aspects.

We compare the performance of the proposed CMPC scheme with two extreme ones; the LFU scheme in which each SBS caches the most popular files, which gives the optimum performance bound in [18], [19], and the Uniform scheme, where the SBSs cache the library files randomly with uniform distribution. It is illustrated in Figs. 2 and 3 that the proposed CMPC scheme has a lower average experienced delay compared to other caching schemes from Uniform to LFU. It should be noted that the MBS and SBSs' average downlink (service) rates are considered equal, to study the network in general, and have a fair comparison.

In Fig. 2, we have shown the effect of the group size $|\bar{S}_g|$, which is the number of the SBSs in each user group, on the values of the average experienced delay. As shown in this figure, by increasing $|\bar{S}_g|$, the values of the average experienced delay increase. Therefore, from the network design perspective, we should have the minimum possible number of the SBSs in each group. In the rest of this section, a network with 4 SBSs with the configuration of $\bar{S}_1 = \{s_1\}, \bar{S}_2 = \{s_2, s_3\}, \bar{S}_3 = \{s_3, s_4\}$ is considered.

Fig. 3 plots the average experienced delay as a function of the ratio of the cache size to the library size, i.e., $\frac{C_n}{F}$, for different schemes. As illustrated in this figure, the most of delay reduction is achieved for the relatively small values of $\frac{C_n}{F}$ and the average response delay slightly changes for large values of $\frac{C_n}{F}$. We have compared the real trace results with the derived equations and simulation results for the IRM traffic. In order to have a comparison among the real trace traffic, the simulation, and analytic results, we need to consider a model for the file popularity distribution for the simulation

and derived equations. We have modeled the file popularity by Zipf distribution [35]. By applying the mean squared error minimization method, 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. We observe that the results achieved under synthetic traffic hold when the cache is fed by real traffic taken from an operational network.

Fig. 4 plots the MBS and SBSs utilization factors of different caching schemes as functions of the ratio of the file size to the average downlink rate, i.e., $\frac{B}{r_0}$. The solid lines show ρ_0 and the others show the maximum values of ρ_n for different schemes. As shown in this figure, the values of ρ_n's are very small in comparison to ρ_0. For example, for the CMPC scheme, the maximum value of ρ_n is about 0.17, where the value of ρ_0 is about 0.95. This figure also shows that ρ_0 of the CMPC scheme is smaller than LFU and Uniform schemes. This means that CMPC reduces the load on the MBS link. In addition, it is illustrated that for large values of $\frac{B}{r_0}$, such as $\frac{B}{r_0} = 5$, LFU and Uniform schemes lead to unstable systems while CMPC scheme ensures the system stability.

Finally, Fig. 5 shows the effect of the bandwidth assignment on the average experienced delay. Solving the optimization problem in Proposition 4 numerically by the MATLAB environment illustrates that the curves are convex, as proven earlier, and the optimum points are shown by red stars. As shown in this figure, by increasing the cache size, the optimum bandwidth assignment (i.e., r_3%) for the proposed CMPC scheme varies from 50.2% to 53.1% in the studied network. In small cache sizes, the average experienced delay slightly changes for different bandwidth assignments. In such cases, the values of ρ_n's are very small compared to ρ_0. Consequently, different bandwidth assignments do not significantly change the values of ρ_n's and $\overline{D}$. By increasing the cache size, the values of ρ_n's and ρ_0 become comparable, and load balancing between MBS and SBSs is done. Therefore, the bandwidth assignment becomes more effective on the average experienced delay in larger cache sizes. We also compare the effect of the bandwidth assignment in the CMPC and LFU schemes. As shown in this figure, the optimum bandwidth assignment for the studied LFU scheme is $r_3 = 61.2$%, and the performance of this scheme is much more sensitive to the bandwidth assignment. For instance, if the bandwidth assigned to r_3 is slightly less than 50%, the system leads to high latency and instability. Therefore, the bandwidth assignment plays an important role in delay minimization, and improper bandwidth assignments lead to large values of the average experienced delay and even an unstable system.

VI. Conclusion

We have proposed a novel analysis framework based on queuing theory and studied the performance of a general and practical structure of HCCNs with overlapping coverage regions and stochastic request arrivals. Based on the framework, we have illustrated the relation between prior network metrics such as the hit probability, link load, and latency in such networks. Then, we have presented the bandwidth assignment problem aiming to minimize the average user-experienced-delay under the stability and cache size constraints. We have also proposed a novel copperative caching scheme in HCCNs that outperforms the existing schemes in terms of delay. The results have been verified through simulations and real trace-driven experiments.

Appendix A
Proof of Proposition 1

We define the random variable $X_{i,n}$ as the number of the requests for file f_i that enter the SBS s_n queue at a given time slot. Therefore, similar to the control unit functionality explained in the proof of Theorem 1 in [16], the average arrival rate of the requests for file f_i entering the SBS s_n queue is given by

$$\lambda_{f_i}^n = \sum_{k=1}^{\infty} P\big(X_{i,n} = k\big) = 1 - P\big(X_{i,n} = 0\big) = p_{serv}(i,n)\, p_{sbs.req}(i,n). \quad \text{(A.1)}$$

Therefore, the average arrival rate at the SBS s_n queue, i.e., λ_n, is evaluated by

$$\lambda_n = \sum_{i=1}^{F} \lambda_{f_i}^n = \sum_{i=1}^{F} p_{serv}(i,n)\, p_{sbs.req}(i,n), \quad \text{(A.2)}$$

and subsequently, if the request arrivals in different user groups are independent, we have

$$\lambda_n = \sum_{i=1}^{F} p_{serv}(i,n)\left(1 - \prod_{g:s_n \in \overline{\mathrm{S}}_g} \big(1 - p_{req}(i,g)\big)\right). \quad \text{(A.3)}$$

In addition, the average service rate at the SBS queue is given by $\mu_n = \frac{r_n}{B}\tau$ [files per time slot]. Consequently, according to (1), (A.2), (A.3), and the definition of ρ_n, (2) and (3) are derived.

Appendix B
Proof of Lemma 1

According to Definitions 1 and 3, we have

$$p_{req}(i,g) = 1 - P\big(N_{\tau,i,g} = 0\big) = 1 - G_{i,g}(\tau, 0). \quad \text{(B.1)}$$

For the renewal traffic model, the Laplace transform of $G_{i,g}(\tau,\xi)$ is obtained from

$$G_{i,g}^*(s,\xi) = \frac{1 - h_g^*(i,s)}{s\big(1 - \xi h_g^*(i,s)\big)}, \quad \text{(B.2)}$$

where $h_g^*(i,s)$ is the Laplace transform of the PDF of the inter-request time distribution for file f_i in the user group M_g [30]. For the IRM traffic model, the distribution of $N_{\tau,i,g}$ is Poisson with mean $\lambda_{req}^g p_i \tau$ which results in (5).

Appendix C
Proof of Proposition 2

According to the HCCN operation framework discussed earlier, the requested contents of the users in M_g that are not served in the caches of the SBSs in $\overline{\mathrm{S}}_g$, are served by the MBS. We define the random variable Y_g^i as the event that at least one request for file f_i from the user group M_g enters

the MBS at a time slot. Consequently, due to the control unit functionality, the requests for file f_i enter the MBS queue with the average arrival rate $\lambda^0_{f_i}$ given by

$$\lambda^0_{f_i} = p\big(\underset{g\in\boldsymbol{G}}{\cup} Y^i_g\big) = \sum_{|J|=1}^{|\boldsymbol{G}|} (-1)^{|J|+1} \sum_{J\subseteq\boldsymbol{G}} p\big(\underset{g\in J}{\cap} Y^i_g\big), \tag{C.1}$$

where the RHS of (C.1) results from the probability of union formula. Moreover, for any given $J \subseteq \boldsymbol{G}$, we have

$$p\big(\underset{g\in J}{\cap} Y^i_g\big) = p_{joint.req}(i,J)\, p_{unserv}(i,J). \tag{C.2}$$

Therefore, by substituting (C.2) in (C.1), the average arrival rate at the MBS queue is obtained from

$$\begin{aligned}\lambda_0 &= \sum_{i=1}^{F} \lambda^0_{f_i} \\ &= \sum_{i=1}^{F} \sum_{|J|=1}^{|\boldsymbol{G}|} (-1)^{|J|+1} \sum_{J\subseteq\boldsymbol{G}} p_{joint.req}(i,J)\, p_{unserv}(i,J).\end{aligned} \tag{C.3}$$

In addition, the average service rate at the MBS queue is given by $\mu_0 = \frac{r_0}{B}\tau$ [files per time slot]. Consequently, according to the definition of ρ_0, (9) is derived.

Appendix D
Proof of Corollary 2

In case that request arrivals in different user groups are independent, we have

$$p_{joint.req}(i,J) = \prod_{g\in J} p_{req}(i,g), \tag{D.1}$$

and in case that requests are independently served at different SBSs, we have

$$p_{unserv}(i,J) = \prod_{n:s_n\in\left(\underset{g\in J}{\cup}\overline{\mathrm{S}}_g\right)} \big(1 - p_{serv}(i,n)\big). \tag{D.2}$$

Moreover, if the requests are almost surely served or not served at the SBSs, then $p_{serv}(i,n)$ only takes 0 and 1. Therefore, if at least one of the SBSs in the SBS groups $\overline{\mathrm{S}}_g : g \in J$, serves a request for file f_i, or in other words, $\exists n : s_n \in \left(\underset{g\in J}{\cup}\overline{\mathrm{S}}_g\right)$ such that $p_{serv}(i,n) = 1$, it means that the request is served in these SBS groups and subsequently, $p_{unserv}(i,J)$ is 0. On the other hand, if no one of the SBSs in the SBS groups $\overline{\mathrm{S}}_g : g \in J$, serves a request, it means that $p_{serv}(i,n) = 0, \forall n : s_n \in \left(\underset{g\in J}{\cup}\overline{\mathrm{S}}_g\right)$. Therefore, the request will not be served in these SBS groups and by definition, $p_{unserv}(i,J)$ will be 1. Consequently, (D.2) is also applied in this case. Therefore, (10) results from Proposition 2, (1), (D.1), and (D.2).

Appendix E
Proof of Proposition 3

For the IRM traffic, we have M/G/1 queue models. According to the Pollaczek-Khinchin (P-K) relation [29], the average delay of delivering a file from the SBS s_n or MBS link is given by

$$dl_n = \frac{1}{\mu_n}\left(1 + \frac{\rho_n\big(1+\upsilon_n^2\big)}{2(1-\rho_n)}\right), \forall n \in \boldsymbol{N}\cup\{0\}. \tag{E.1}$$

Combining (11)-(13) and (E.1) results in (14).

Appendix F
Proof of Proposition 4

The optimization problem formulation in Proposition 4 results from Propositions 1 and 3. Noting that $r_n, \forall n \in \boldsymbol{N}$, are unknown variables, the partial derivatives of the objective function are given by

$$\frac{\partial\overline{D}}{\partial r_n} = \frac{1}{|\boldsymbol{G}|}\sum_{g\in\boldsymbol{G}} \frac{p_{serv}(g)}{|\overline{\mathrm{S}}_g|}\left(-\frac{B}{\tau r_n^2} - \frac{\frac{B^2}{\tau^2}\lambda_n\big(1+\upsilon_n^2\big)(4r_n - 2\frac{B}{\tau}\lambda_n)}{4r_n^2(r_n - \frac{B}{\tau}\lambda_n)}\right), \tag{F.1}$$

$$\begin{aligned}\frac{\partial^2\overline{D}}{\partial r_n r_m} = 0, \forall m \neq n, \qquad \frac{\partial^2\overline{D}}{\partial r_n^2} &= \frac{1}{|\boldsymbol{G}|}\sum_{g\in\boldsymbol{G}} \frac{p_{serv}(g)}{|\overline{\mathrm{S}}_g|} \\ &\left(\frac{2B}{\tau r_n^3} + \frac{\frac{B^2}{\tau^2}\lambda_n\big(1+\upsilon_n^2\big)\big(\frac{B^2}{\tau^2}\lambda_n^2 + 3r_n(r_n - \frac{B}{\tau}\lambda_n)\big)}{r_n^3(r_n - \frac{B}{\tau}\lambda_n)^3}\right).\end{aligned} \tag{F.2}$$

Considering the stability constraint in (19), we have $\frac{\partial^2\overline{D}}{\partial r_n^2} \geq 0$. Therefore, the objective function is convex. Sinc the constraints are linear, it is a convex optimization problem [33].

References

[1] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective,"*IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 118-127, Jun. 2014.

[2] J. Erman, A. Gerber, and M.T. Hajiaghayi, "To Cache or Not to Cache-The 3G Case," *IEEE Internet Computing*, vol. 15, no. 2, pp. 27-34, Apr. 2011.

[3] B.A. Ramanan, L.M. Drabeck, M. Haner, N. Nithi, T.E. Klein, and C. Sawkar, "Cacheability Analysis of HTTP traffic in an Operational LTE Network," *Wireless Telecom. Sympos.*, pp. 1-8, Jul. 2013.

[4] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting Caching and Multicast for 5G Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995-3007, Apr. 2016.

[5] X. Wang, M. Chen, T. Taleb, A. Ksentini, and 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, Feb. 2014.

[6] S. Krishnan, M. Afshang, and H. S. Dhillon, "Effect of Retransmissions on Optimal Caching in Cache-Enabled Small Cell Networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11383-11387, Dec. 2017.

[7] Y. Zhou, F. R. Yu, J. Chen, and Y. Kuo, "Resource Allocation for Information-Centric Virtualized Heterogeneous Networks With In-Network Caching and Mobile Edge Computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11339-11351, Dec. 2017.

[8] 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, Apr. 2013.

[9] 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, Oct. 2015.

[10] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint Caching, Routing, and Channel Assignment for Collaborative Small-Cell Cellular Networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275 - 2284, Jun. 2016.

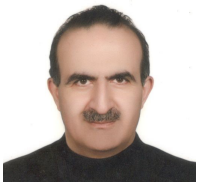
[11] B. Serbetci and J. Goseling, "On Optimal Geographical Caching in Heterogeneous Cellular Networks," *in Proc. IEEE Wireless Commun. Net. Conf. (WCNC)*, Mar. 2017.

[12] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Joint Caching and Base Station Activation for Green Heterogeneous Cellular Networks," *in Proc. IEEE Int. Conf. Commun (ICC)*, Jun. 2015.

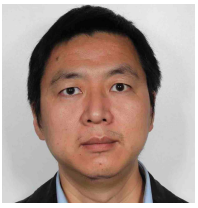
[13] J. Zhang, X. Zhang, M. A. Imran, B. Evans, and W. Wang, "Energy Efficiency Analysis of Heterogeneous Cache-enabled 5G Hyper Cellular Networks," *in Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dec. 2016.

[14] D. Liu and C. Yang, "Cache-enabled Heterogeneous Cellular Networks: Comparison and Tradeoffs," *in Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016.
[15] F. Rezaei, A. Momeni, and B. H. Khalaj, "Delay analysis of network coding in multicast networks with Markovian arrival processes: A practical framework in cache-enabled networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7577-7584, 2018.
[16] 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, Jan. 2016.
[17] F. Rezaei, B. H. Khalaj, M. Xiao, and M. Skoglund, "Delay and Stability Analysis of Caching in Heterogeneous Cellular Networks," *in Proc. IEEE Int. Conf. Telecom. (ICT)*, May 2016.
[18] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on Cache-enabled Wireless Heterogeneous Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp.131-145, Jan. 2016.
[19] T. D. Tran, T. D. Hoang, and L. B. Le, "Caching for Heterogeneous Small-Cell Networks With Bandwidth Allocation and Caching-Aware BS Association," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 49-52, Feb. 2019.
[20] B. Zhou, Y. Cui, and M. Tao, "Stochastic Content-Centric Multicast Scheduling for Cache-Enabled Heterogeneous Cellular Networks," *IEEE Trans. Wireless. Commun.*, vol. 15, no. 9, Sep. 2016.
[21] W. Jiang, G. Feng, and S. Qin, "Optimal Cooperative Content Caching and Delivery Policy for Heterogeneous Cellular Networks," *IEEE Trans. Mob. Comput.*, vol. 16, no. 5, pp. 1382-1393, May 2017.
[22] B. N. Bharath, K. G. Nagananda, and H. Vincent Poor, "A Learning-Based Approach to Caching in Heterogeneous Small Cell Networks," *IEEE Trans. Commun.*, vol. 64, no.4, pp. 1674-1686, Apr. 2016.
[23] M. Ji, G. Caire, and A. F. Molisch, "The Throughput-Outage Tradeoff of Wireless One-Hop Caching Networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833-6859, Dec. 2015.
[24] S. A. A. Siahpoosh and F. Rezaei, "A Mobility-Aware Caching Scheme in Heterogeneous Cellular Networks," *in Proc. IEEE Int. Comput. Conf. (CSICC)*, Mar. 2021.
[25] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation Algorithms for Mobile Data Caching in Small Cell Networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665-3677, Oct. 2014.
[26] 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)*, Apr. 2014.
[27] S. Podlipnig and L. Boszormenyi, "A survey of Web cache replacement strategies," *ACM Computing Surveys*, vol. 35, no. 4, Dec. 2003.
[28] E. Rosensweig, J. Kurose, and D. Towsley, "Approximate models for general cache networks," *in Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Mar. 2010.
[29] L. Kleinrock, *Queueing Systems*, vol. 1, Hoboken, NJ, USA: Wiley, 1975.
[30] D. R. Cox, *Renewal Theory*, London. U.K.: Methuen,1962.
[31] Y. E. Sagduyuand and A. Ephremides, "On broadcast stability of queue-based dynamic network coding over erasure channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5463-5478, Dec. 2009.
[32] W. Cheung, T. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, access control in two-tier femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 561-574, Apr. 2012.
[33] S. Boyd and L. Vandenberghe, *Convex Optimization*, New York, USA, Cambridge University Press, 2004.
[34] 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, Mar. 2009.
[35] L. Shi1, Z. Gu, L.Wei, and Y. Shi, "An applicative study of Zipf's law on web cache," *Int. J. Inf. Technol.*, vol. 12, no. 4, pp. 49-58, Jan. 2006.

Fatemeh Rezaei received the B.Sc., M.Sc., and PhD degrees in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2010, 2012, and 2017, respectively. She spent a sabbatical stay at KTH University, Stockholm, Sweden, in 2015. She is currently an Assistant Professor with the department of computer engineering, K. N. Toosi University of Technology, Tehran, Iran. Her current research interests include caching in wireless networks and edge computing.



Babak Hossein Khalaj received his B.Sc. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 1989, and M.Sc. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, USA, in 1993 and 1996, respectively. Since 1999, he has been a Senior Consultant in the area of data communications, and from 2006 to 2007, a Visiting Professor with CEIT, San Sebastian, Spain. He has coauthored many papers in signal processing and digital communications and holds four U.S. patents. He was the recipient of the Alexander von Humboldt Fellowship from 2007 to 2008 and Nokia Visiting Professor Scholarship in 2018.



Ming Xiao received Bachelor and Master degrees in Engineering from the University of Electronic Science and Technology of China, ChengDu in 1997 and 2002, respectively. He received Ph.D degree from Chalmers University of technology, Sweden in November 2007. From November 2007 to now, he has been in the department of information science and engineering, school of electrical engineering and computer science, Royal Institute of Technology, Sweden, where he is currently an Associate Professor. He was an Editor for IEEE Transactions on Communications (2012-2017), IEEE Communications Letters (Senior Editor Since January 2015) and IEEE Wireless Communications Letters (2012-2016), and has been an Editor for IEEE Transactions on Wireless Communications since 2018. He has been an area editor for IEEE Open Journal of the Communication Society since 2019.

Mikael Skoglund received the Ph.D. degree in 1997 from Chalmers University of Technology, Sweden. In 1997, he joined the Royal Institute of Technology (KTH), Stockholm, Sweden, where he was appointed to the Chair in Communication Theory in 2003. At KTH, he heads the Division of Information Science and Engineering, and the Department of Intelligent Systems. He has authored and co-authored more than 185 journal and 400 conference papers. Dr. Skoglund is a Fellow of the IEEE. During 2003–08 he was an associate editor for the IEEE Transactions on Communications. During 2008–12 he was on the editorial board for the IEEE Transactions on Information Theory and starting in the Fall of 2021 he joined it once more. He has served on numerous technical program committees for IEEE sponsored conferences, he was general co-chair for IEEE ITW 2019, and he will serve as TPC co-chair for IEEE ISIT 2022.