A Mobility-Aware Caching Scheme in Heterogeneous Cellular Networks

Seyyed Amir Ahmad Siahpoosh
Computer Engineering Department
K. N. Toosi University of Technology
Tehran, Iran
s.a.a.siahpoosh@email.kntu.ac.ir

Fatemeh Rezaei
Computer Engineering Department
K. N. Toosi University of Technology
Tehran, Iran
frezaei@kntu.ac.ir

Abstract— **User mobility is a challenging problem in heterogeneous cellular networks. In this paper, we try to turn the mobility challenge into an opportunity to reduce latency. We first define a model of an urban cellular network in which mobile users can move between different small cells. Then, by introducing a scheme called Cooperative LRU, we use user mobility to reduce the file download delays. In this method, the requested file which is cached in the current user's cell is also cached in two adjacent cells. This means that the caching scheme in the current cell is reactive and in two other adjacent cells is proactive. Finally, we have a comparison between traditional methods and the introduced method and we examine the effect of different network and kinetic parameters on reducing latency.**

Keywords— caching, mobility, cellular networks, heterogeneous networks, urban networks.

I. INTRODUCTION

In recent years, due to the fast development of smart phone and tablet technologies, mobile data traffic has been growing exponentially [1]. In 2019, IP video traffic accounts for 80% of total traffic, and 81% of all connected devices are mobile. Furthermore, by 2022 mobile video will generate nearly four-fifths of mobile data traffic [2]. It should be noted that a large portion of backhaul traffic is contributed by transmitting duplicate popular data to multiple users [1]. On the other hand, technologies such as Internet of Things, connected vehicles, smart grid, and augmented reality, have new requirements like low delay, high data rate, high Quality of Service (QoS) and low power [3]. To address these problems, various solutions have been presented. Mobile Edge Computing (MEC), which deploys cloud servers in base stations, is a promising solution for the problem since the computation capability is closer to mobile devices [3]. Device to Device (D2D) communications, utilizing the same spectrum as a cellular user, is regarded as a new paradigm with great potential for supporting proximity-based applications [4]. Among the potential solutions, caching techniques have attracted significant attention since they can effectively reduce the backhaul traffic by eliminating duplicate data transmission that carries popular content [1].

During the past several decades, mobile cellular networks have been evolving steadily and significantly from the 1st generation (1G) voice-only systems to current 4th generation (4G) all-IP based LTE-Advanced networks [3]. The 5th Generation of mobile networks (5G) presents solutions that go beyond performance enhancements of the radio link, making new enablers available to network operators and service providers allowing the network to become more flexible and to dynamically address the changing needs of running services [2]. The heterogeneous network architecture in 5G networks with the dense deployment of small-cell base stations (SBSs) in coexistence with the macro-cell base stations (MBSs) provides an important solution to better satisfy the ever-growing connected devices and mobile traffic [5]. By bringing the base stations closer to the devices in small cell networks, the spectral efficiency can be significantly enhanced due to increased spatial reuse [6]. Caching in such networks has been studied in some recent works from different perspectives [6-10]. There are some challenges to implement these architectures and technologies in practice. One of the most important issues is how to consider the effects of the users' mobility on caching techniques in such networks.

The paper is organized as follows. Section II discusses the related works. The system model is presented in Section III. In Section IV, the performance evaluation through numerical results is presented. Finally, Section V concludes the paper.

II. RELATED WORK

There are some recent works considering the mobility of users in cache-enabled networks. The authors in [2] considered caching for fast vehicles such as trains. They used network virtualization where the Software-Defined Network (SDN) controller knows the data flow and predicts the user mobility pattern and sends the next chunk of video to the next base station. A mobility-aware content placement model was proposed in [11]. The goal of this paper is to maximize the cache hit ratio and the energy consumption for content delivery and give the optimal transmission power of SBSs and mobile devices. Since the optimization problem is NP-hard, the authors solved the sub-optimal problem with a greedy algorithm. This work assumed pairwise contact process is independent Poisson Process. However, it was shown that the hypothesis hexagonal grid and pairwise contact Poisson process for the location model are not realistic [12]. The authors in [12] introduced two subsets of Gibbs distribution for modeling the node locations and illustrated that the Strauss process for the inhibitive deployment and the Geyer saturation process for the clustered deployment is better assumptions than the Poisson process. They also used Voronoi cell area distributions to show that variations of the hexagonal grid do not accurately model the coverage cell size.

In [13], proactive caching was considered and the speeds of the vehicles were assumed to be independent and identically distributed (i.i.d) and generated by the truncated

978-0-7381-2628-9/21/$31.00 ©2021 IEEE

Gaussian distribution. The number of the arrived vehicles for entering each entrance during the defined period followed the Poisson process. They introduced Mobility-aware Proactive Caching based on Federated learning (MPCF) algorithm for caching the popular contents and employed a Context-aware Adversarial Auto Encoder (C-AAE) to predict the highly dynamic content popularity. The authors in [14] defined the effective D2D coverage area of a helper user with its velocity and time. They used the hit ratio and communication cost as the performance metrics. In [15], the area was partitioned into hot regions and the caching nodes in hot regions were chosen based on the sojourn time. They used a method for predicting the next visiting area and the trajectory history, hit ratio and delay were considered as the metrics.

III. System Model

A. Network Architecture

We consider a heterogeneous cellular network in an urban area consisting of a macro-cell and several small cells, each of which overlaps with its two neighbors as shown in Fig. 1. The SBSs download rates are considered greater than the MBS, according to the non-standalone deployment of 5G networks [16]. In order to model the users' mobility, we consider a square at the center of each small cell, where M mobile users (MUs) move there. As shown in Fig. 1, each square consists of two entrances and two exits that allow moving between the squares. The movement of the users in the squares is one-way and clockwise.

In the beginning, there are some mobile users in the area, whose initial locations are randomly assigned. The mobile users randomly exit one of the exits and enter the other area, and this process continues until the end. The speeds of the MUs are assumed to be independent and identically distributed (i.i.d.), denoted by a set $U = \{U_1, U_2, U_3, \ldots, U_m\}$. They are generated by a truncated Gaussian distribution with a specific mean. This assumption has also been widely used in many state-of-the-art works of vehicular networks [13]. Compared to the normal Gaussian distribution or a fixed speed, the truncated Gaussian distribution is more feasible for modeling vehicles' speed because it limits the scope of vehicles speed to a certain range.

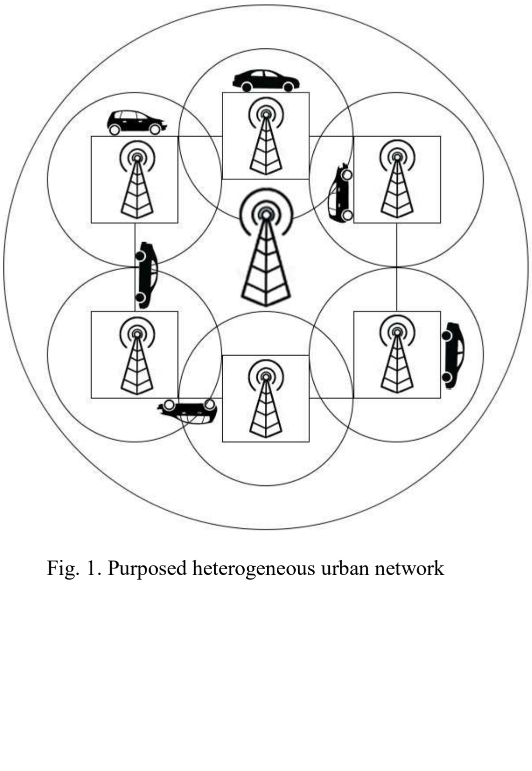

Fig. 1. Purposed heterogeneous urban network

TABLE I. Key Notations

Notation	Semantic
M	Number of Users
C	Cache Size
B	Size of Files
U_i	User Speed
μ	Mean of Gaussian Distribution

The user requests are drawn from a specific same-size file library $\boldsymbol{F}= \{f_i : i = 1, \ldots, F\}$ of size B bits, and the SBSs are capable of caching C whole files (i.e., CB bits). In order to model statistical request arrivals, we consider the IRM traffic model for the stream of the requests, which is based on the following fundamental assumptions [7]: i) users request items from a fixed library of F files; ii) the probability that a request is for file f_i is constant (i.e., the file popularity does not vary over time), and is also independent of all past requests, generating an i.i.d. sequence of requests.

Table I shows key notations adopted for the system model and subsequent analysis presented in this paper.

B. Proposed Cooperative LRU Method

In this part, we present a mobility-aware caching scheme in order to reduce file transmission delay using user mobility. In our proposed scheme, which is called Cooperative LRU, each user sends his request to the base stations. After receiving the request, if the requested file hits in the SBS cache, the SBS will respond to the request and start sending the file; but if the file does not exist in the SBS cache, the MBS will respond to the request and send the file to the corresponding SBS and its two neighbors, simultaneously. The SBSs cache the coming files using the LRU method [8]. Using the proposed scheme, if the user leaves the current small cell while receiving the file, he can continue downloading the rest of the file from the new SBS. Moreover, if the other users request the same file in the neighboring SBSs, they can download the file from the SBSs as long as it is available in their caches.

According to the greater downloading rates from the SBSs, the proposed scheme reduces the file download latency. The performance improvement depends on system parameters such as user speed, file size and cache size. The larger the cache, the more likely it is to receive the file from the SBSs, and the efficiency of the proposed scheme would be improved. Also, the larger the file size, the more likely it is to leave the requested area while receiving the file, and the method will be more effective. User speed has a similar effect. If the user speed is fast enough so that the files are not completely downloaded before leaving a small cell, the proposed scheme will significantly improve the performance.

In the next section, we examine the effect of the proposed scheme on reducing the file download latency and compare it with conventional caching. We also investigate the effect of various parameters in improving the performance of the proposed scheme.

IV. Performance Evaluation and Numerical Results

To evaluate the performance of the proposed scheme, we simulate the network using NS2. The radius of each small cell is 1 km and the radius of the macro cell is 3 km. Also, the download rate from the SBSs is considered 1 Gbps and from the MBS is 100 Mbps. We use the IRM model to simulate network traffic, similar to the model used in [8]. According to the real traffic in a campus network measurement on YouTube traffic in 2008 [17], a total 123.3k requests for 78.9k videos has been considered. We consider M = 18 MUs in the macro cell. At the beginning of the simulation in each SBS, there are 3 users whose locations are determined randomly, and then, these users move between different areas. We compare our proposed scheme with the typical LRU method [7]. In the typical LRU method, caching is done separately in each small cell and the SBSs have no cooperation with each other. In fact, in the typical LRU method, we do not have proactive caching and reactive caching occurs in only one small cell. In the following, we see the simulation results with different values of the parameters.

In Fig. 2, the delay changes in terms of the cache size is illustrated. In this scenario, the file size is considered to be 1 Gb and the users' speed follows the truncated Gaussian distribution with μ=15, min=5 and Max=25 m/s. The orange line represents the typical LRU caching method and the blue line shows the results of the proposed scheme. It is clear that in both methods the delay decreases with increasing the cache size, but in general, the proposed method works better than conventional caching. It should be noted that the proposed method is more superior in small cache sizes than the traditional method; because the larger the cache, the more files are stored in it and mobility will not have much effect, which requires more cost.

Fig. 3 shows the effect of the average user speed on the download latency. In this scenario, the cache size is considered 20% of the total files and the size of each file is 100 Mb. It is observed that by increasing the average speed of the users, the delay of downloading the files in the proposed method significantly decreases, because the probability that the requesting user leaves the initial small cell becomes higher; consequently, by entering the new small cell, the content is already cached and available. Another noteworthy point is that increasing the speed of the users to some extent reduces the delay, and then the slope of delay reduction decreases. On the other hand, it is clear that the speed of the users does not affect the performance of conventional LRU caching.

Fig. 4 and Fig. 5 show the effect of the file size on delay. In Fig. 4, the cache size is 20% of the total files and in Fig. 5, the cache size is 40% of the total files. The mean users' speed in both cases is 15 m/s. According to the diagrams, in the case of small files and high download rates, the proposed method has a little delay reduction compared to conventional LRU caching. Because downloading the small files usually ends in the initial small cell, as explained earlier. But in the case of large files, it can be seen that the proposed method has a dramatic effect in reducing delay.

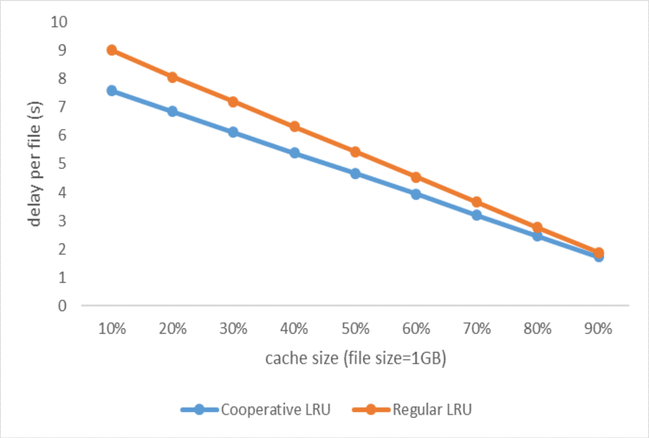

Fig. 2. Delay vs. cache size (average speed=15 m/s, file size=1Gb)

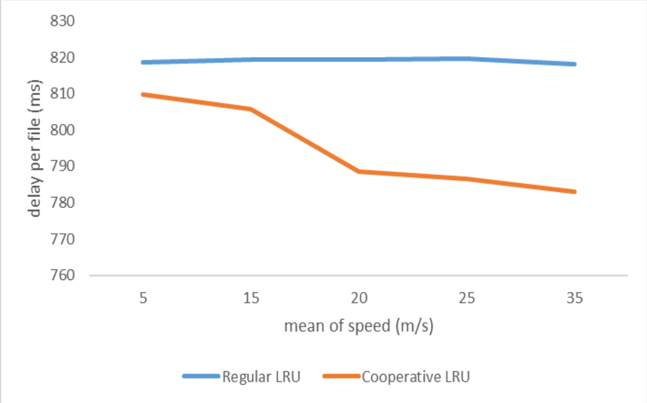

Fig. 3. Delay vs. mean speed (cache size=20%, file size=100Mb)

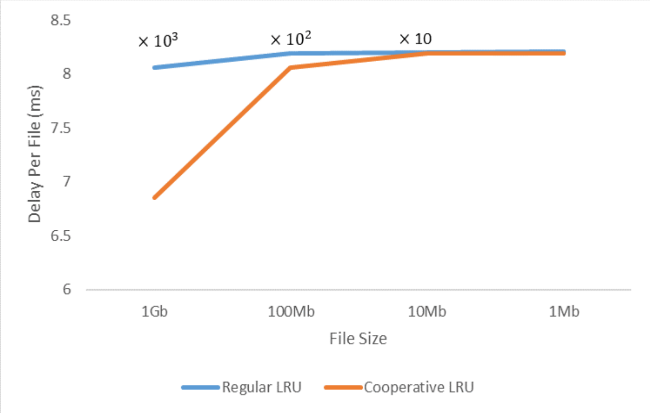

Fig. 4. Delay vs. file size (cache size=20%, average speed=15 m/s)

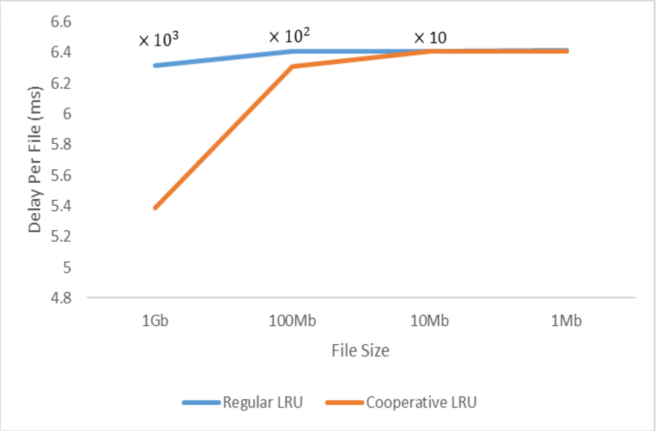

Fig. 5. Delay vs. file size (cache size=40%, average speed=15 m/s)

V. Conclusion

In this paper, we study the effects of mobility on the delay of receiving a file in a heterogeneous urban network. We introduced a method called Cooperative LRU caching, which can be used to reduce latency in a heterogeneous urban network by considering user mobility. We also observed the effect of some motion parameters such as speed and network parameters such as file size and cache size in improving the performance of the introduced model. For future work, the network model can be expanded and different movement models can be studied according to the real-world movement patterns.

References

[1] L. Li, G. Zhao, and R. S. Blum, "A Survey of Caching Techniques in Cellular Networks: Research Issues and Challenges in Content Placement and Delivery Strategies*", IEEE Commun. Surveys and Tutorials,* vol. 20, no. 3, pp. 1710-1732, 2018**.**

[2] R. Silva, D. Santos, D. Corujo, R. L. Aguiar, S. Figueriedo, and B. Parreira, "Mobility-Optimized Dynamic Content Placement for Fast Vehicles in 5G Networks", *IEEE Sympos. PIMRC*, Istanbul, 8-11 Spt. 2019.

[3] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications", *IEEE Access,* vol. 5, pp. 6757-6779, 2017.

[4] M. Ahmed, Y. Li, M. Waqas, M. Sheraz, D. Jin, and Z. Han, "A Survey on Socially Aware Device-to-Device Communications", *IEEE Commun. Surveys and Tutorials,* vol. 20, no.3, pp. 2169-2197, 2018.

[5] T. D. Tran, T. Duc Hoang, and L. B. Le, "Caching for Heterogeneous Small-Cell Networks With Bandwidth Allocation and Caching-Aware BS Association", *IEEE Wireless Commun. Letters,* vol. 8, no. 1, pp. 49-52, 2018.

[6] 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.

[7] 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.

[8] F. Rezaei, B. H. Khalaj, M. Xiao, and M. Skoglund, "Delay and Stability Analysis of Caching in Heterogeneous Cellular Networks," *IEEE Int. Conf. Telecom. (ICT 16),* May 2016.

[9] 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.

[10] 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. Vehicular Technol.,* vol. 67, no. 8, pp. 7577-7584, 2018.

[11] M. Chen, Y. Hao, L. Hu, K. Huang, and V. K. N. Lau, "Green and Mobility-Aware Caching in 5G Networks", *IEEE Trans. Wireless Commun.,* vol. 16, no.12, pp. 8347-8361, 2017.

[12] D. B. Taylor, H. S. Dhillon, T. D. Novlan, and J. G. Andrews, "Pairwise interaction processes for modeling cellular network topology", *IEEE Global Commun. Conf.,* Anaheim, Dec. 2012.

[13] Zh. Yu, J. Hu, G. Min, Zh. Zhao, W. Miao, and M. S. Hossain, "Mobility-Aware Proactive Edge Caching for Connected Vehicles Using Federated Learning", *IEEE Trans. Intelligent Transport. Systems*, Early Access, 2020.

[14] S. Kim, E. Go, Y. S. Song, H. J. Cho, M. Rim, and C. G. Kang, "A Study on D2D Caching Systems with Mobile Helpers", *Int. Conf. Ubiquitous and Future Networks (ICUFN*), Prague, 2018.

[15] L. Yao, A. Chen, J. Deng, J. Wang, and G. Wu, "A Cooperative Caching Scheme Based on Mobility Prediction in Vehicular Content Centric Networks", *IEEE Trans. Vehicular Technol.,* vol. 67, no.6, pp. 5435-5444, 2018.

[16] D. S. Michalopoulos, I. Viering, and L. Du, "User-Plane Multi-Connectivity Aspects in 5G", *IEEE Int. Conf. Telecom., (ICT 16),* 2016.

[17] 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. Telecom.,* Netw., vol. 53 no. 4, pp. 501–514, Mar. 2009.