Mobility-Aware Caching in D2D Networks

Vikas B Wasade [1], Nilesh Bodne[2]

[1]Department of Electronics (Comm. Engineering), R.T.M. Nagpur University
[2]Department of Electronics (Comm. Engineering), R.T.M. Nagpur University

ABSTRACT

In this paper, we propose a novel policy for device caching that facilitates popular content exchange through highrate device-to-device (D2D) millimeter-wave (mmWave) communication.The D2D aware caching (DAC) policy splits the cacheable content into two content groups and distributes it randomly to the user equipment devices (UEs), with the goal to enable D2D connections. By exploiting the high-bandwidth availability and the directionality of mmWaves, we ensure high rates for the D2D transmissions, while mitigating the co-channel interference that limits the D2D-communication potentials in the sub-6 GHz bands. Furthermore, based on a stochasticgeometry approach for the modeling of the network topology, we analytically derive the offloading gain that is achieved by the proposed policy and the distribution of the content retrieval delay considering both half- and full-duplex mode for the D2D communication. The accuracy of the proposed analytical framework is validated through Monte-Carlo simulations. In addition, for a wide range of a content popularity indicator the results show that the proposed policy achieves higher offloading and lower content-retrieval delays than existing state-of-the-art approaches.

Keyword - *Caching device-to-device communications, human mobility, matroid constraint, submodular function.*

1. Introduction

Over the last few years, the proliferation of mobile devices connected to the Internet, such as smartphones and tablets, has led to an unprecedented increase in wireless traffic that is expected to grow with an annual rate of 53% until 2020. To satisfy this growth, a goal has been set for the 5th generation (5G) of mobile networks to improve the capacity of current networks by a factor of 1000. While traditional approaches improve the area spectral efficiency of the network through, e.g., cell densification, transmission in the millimeterwave (mmWave) band, and massive MIMO, studies have highlighted the repetitive pattern of user content requests, suggesting more efficient ways to serve them.

With proactive caching, popular content is stored inside the network during off-peak hours (e.g., at night), so that it can be served locally during peak hours. Two methods are distinguished in the literature: i) edge caching when the content is stored at helper nodes, such as small-cell basestations (BSs), and ii) device caching when the content is stored at the user equipment devices (UEs). While edge caching alleviates the backhaul constraint of the small-cells by reducing the transmissions from the core network, device caching offloads the BSs by reducing the cellular transmissions, which increases the rates of the active cellular UEs and reduces the dynamic energy consumption of the BSs. The UEs also experience lower delays since the cached content is served instantaneously or through D2D communication from the local device caches.

The performance of the mmWave bands in wireless communication has been investigated in the literature for both outdoor and indoor environments, especially for the frequencies of 28 and 73 GHz that exhibit small atmospheric absorption. According to these works, the coverage probability and the average rate can be enhanced with dense mmWave deployments when highly-directional antennas are employed at both the BSs and the UEs. MmWave systems further tend to be noise-limited due to the high bandwidth and the directionality of communication. Recently, several works have conducted system-level analyses of mmWave networks with stochastic geometry, where the positions of the BSs and the UEs are modeled according to homogeneous Poisson point processes (PPPs). This modeling has gained recognition due to its tractability.

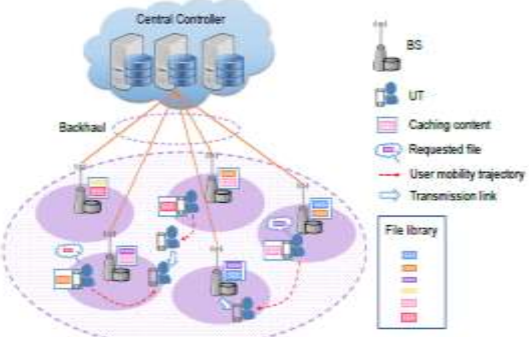


Fig -1: A sample cache

2. Literature Survey:

Compared with caching at BSs, caching at mobile devices provides new features and unique advantages. First, the aggregate caching capacity grows with the number of devices in the network, and thus the benefit of device caching improves as the network size increases. Second, by caching popular content at devices, mobile users may acquire required files from close user devices via D2D communications, rather than through the BS. This will significantly reduce the mobile traffic on the backbone network and alleviate the heavy burden on the backhaul links. However, mobile caching also faces some new challenges. Specifically, not only the users, but also the caching helpers are moving over time, which brings additional difficulties in the caching design. There have been lots of efforts on D2D caching networks while assuming fixed network topologies. Considering a single cell scenario, it was shown in that D2D caching outperforms other schemes, including the conventional unicasting scheme, the harmonic broadcasting scheme, and the coded multicast scheme. Assuming that each device can store one file, Golrezaei *et al.* analyzed the scaling behavior of the number of active links in a D2D caching network as the number of mobile devices increases .It was found that the concentration of the file request distribution affects the scaling laws, and three concentration regimes were identified. In the outage-throughput tradeoff in D2D caching networks was investigated and optimal scaling laws of per-user throughput were derived as the numbers of mobile devices and files in the library grow under a simple uncoded protocol. Meanwhile, the case using the coded delivery Scheme was investigated in. There are some preliminary studies considering user mobility. In Poularakis *et al.* studied a femto-caching network with user mobility. A Markov chain model was adopted to represent which helper, i.e., a particular femtocell BS, was accessed by a specific user in different time slots. However, in D2D caching networks, such a model cannot be adopted, since there is no fixed caching helper, and all the mobile users may move over time. The effect of user mobility on D2D caching was investigated via simulations in which showed that user mobility does not have a significant impact on a random caching scheme. However, such a caching scheme failed to take advantage of the user mobility pattern. In it showed that user mobility has positive effect on D2D caching.In Lan et al. considered the case where mobile users can update caching content based on the file requirement and user mobility. However, it was assumed that one complete file can be transmitted via any D2D link when two users contact, which is not practical, considering the limited communication time and transmission rate. More recently, several design methodologies for mobility-aware caching were proposed in but more thorough investigations will be needed, especially on practical implementations.

3. Device Caching Model

We assume that the UEs request content from a library of L files of equal size _f ile and that their requests follow the Zipf distribution. According to this model, after ranking the files with decreasing popularity, the probability qi of a UE requesting the i-th ranked file is given by

$$q_i = \frac{i^{-\xi}}{\sum_{j=1}^{L} j^{-\xi}}, \ 1 \leq i \leq L, \ \xi \geq 0,$$

where qi is the popularity exponent of the Zipf distribution.This parameter characterizes the skewness of the popularity distribution and depends on the content1 type, (e.g., webpages, video, audio, etc.). In device caching, every UE retains a cache of K files, where K << L, so that when a cached content is requested, it is retrieved locally with negligible delay instead of a cellular transmission. This event is called a cache hit and its probability is called the hit probability, which is denoted by h and given by

$$h = \sum_{i \in C} q_i,$$

where C represents the cached contents of a UE, as determined by the caching policy.

The MPC policy is a widely considered caching scheme that stores the K most popular contents from the library of L files in every UE. This content placement maximizes the hit probability, which is given by

$$h_{mpc} = \sum_{i=1}^{K} q_i = \frac{\sum_{i=1}^{K} i^{-\xi}}{\sum_{j=1}^{L} j^{-\xi}}.$$

3.1 Proposed DAC Policy-

Although the MPC policy maximizes the hit probability, it precludes content exchange among the UEs since all of them store the same files. In contrast, a policy that diversifies the content among the UEs enables the content exchange through D2D communication, resulting in higher offloading. Furthermore, thanks to the high D2D rate and the enhancement in the cellular rate due to the offloading, the considered policy may also improve the content retrieval delay, despite its lowe hit probability compared with the MPC policy.

Based on this intuition, in the proposed DAC policy, the 2K most popular contents of the library of L files are partitioned into two non-overlapping groups of K files, denoted by groups A and B, and are distributed randomly to the UEs, which are characterized as UEs A and B respectively. When a UE A is close to a UE B, the network may pair them to enable content exchange through D2D communication. Denoting by hA and hB the hit probabilities of the two UE types, three possibilities exist when a paired UE A requests content:

- The content is retrieved through a cache hit from the local cache of UE A with probability hA.
- The content is retrieved through a D2D transmission from the cache of UE B with probability hB.
- The content is retrieved through a cellular transmission from the associated BS of UE A with probability 1 -hA - hB.

The above cases are defined accordingly for UE B. In Proposition 1 that follows, we formally prove that the probability of content exchange for both paired UEs are maximized with the content assignment of the DAC policy.

When the caches of UEs A and B are non-overlapping, the hit probabilities of two paired UEs coincide with their content exchange probabilities, i.e., eA = hB and eB = hA, hence, the DAC policy also maximizes hA and hB over all possible 2K partitions in the sense of Proposition 12. The 2K most popular contents can be further partitioned in multiple ways, but one that equalizes hA and hB is chosen for fairness considerations. Although exact equalization is not possible due to the discrete nature of the Zipf distribution, the partition that minimizes the difference jhA □ hBj can be found. Considering that this difference is expected to be negligible for sufficiently high values of K, hA and hB can be expressed as

$$h_A \approx h_B \approx h_{dac} = \frac{1}{2}\sum_{i=1}^{2K} q_i.$$

Finally, since two paired UEs may want to simultaneously exchange content, with probability h2 dac, we consider two cases for the DAC policy: i) an HD version, denoted by HDDAC, where the UEs exchange contents with two sequential HD transmissions, and ii) an FD version, denoted by FDDAC, where the UEs exchange contents simultaneously with one FD transmission. Although the FD-DAC policy increases the frequency reuse of the D2D transmissions compared with the HD-DAC policy, it also introduces self-interference (SI) at the UEs that operate in FD mode and increases the D2D cochannel interference. It therefore raises interesting questions regarding the impact of FD communication on the rate performance, especially in a mmWave system where the co-channel interference is naturally mitigated by the directionality.

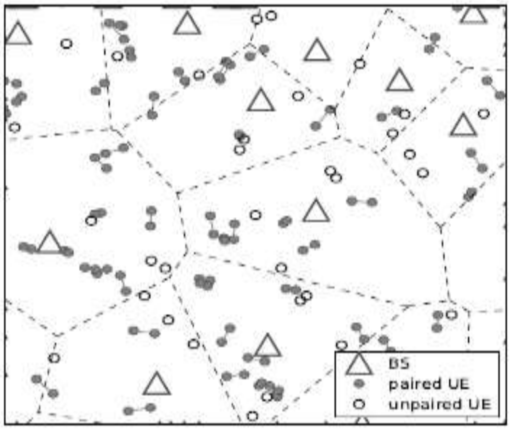


Fig-2: A network snapshot in a rectangle dimensions

3.2 System Model

In Mobility-Aware Caching in D2D Networks we take advantage of the user mobility pattern by the inter-contact times between different users,and propose a mobility-aware caching placement strategy to maximize the data offloading ratio which is defined as the percentage of the requested data that can be delivered via D2D links rather than through base stations (BSs).

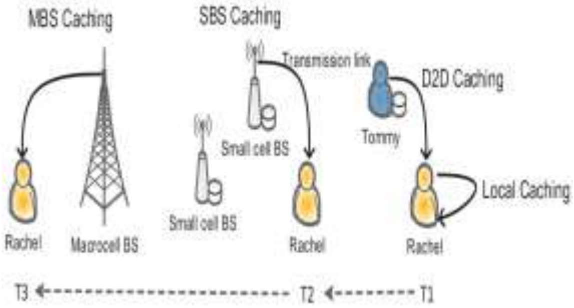


Fig – 3:A system model network

Most previous investigations on wireless caching networks assumed fixed network topologies. However, user mobility is an intrinsic feature of wireless networks, which changes the network topologies over time. Thus, it is critical to take the user mobility pattern into account. On the other hand, user mobility can also be a useful feature to exploit, as it will increase the communication opportunities of moving users. Mobility-aware design has been proved to be an effective approach to deal with lots of problems in wireless networks. For example, exploiting user mobility helps improve capacity in ad hoc networks and reduce the probability of failed file delivery in femto-caching networks. In this paper, we will propose an effective mobility-aware caching strategy in device-to-device (D2D) caching networks to offload traffic from the cellular network.

3.3 Resource Allocation and Scheduling

We focus on the downlink of the cellular system, which is isolated from the uplink through frequency division depluxing (FDD), since the uplink performance is not relevant for the considered caching scenario. We further consider an inband overlay scheme for D2D communication, where a fraction d2d of the overall downlink spectrum BW is reserved for the D2D traffic, justified by the availability of spectrum in the mmWave band. Regarding the scheduling scheme, we consider TDMA scheduling for the active cellular UEs, which is suited to mmWave communication, and uncoordinated D2D comnunication for the D2D UEs, relying on the directionality of the mmWave transmissions for the interference mitigation.

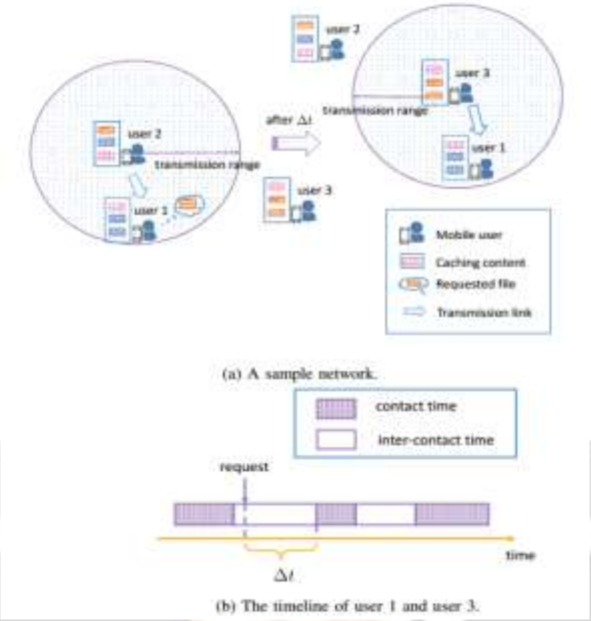
3.4 Offloading

Mobile data offloading, is the use of complementary network technologies for delivering data originally targeted for cellular networks. Offloading reduces the amount of data being carried on the cellular bands, freeing bandwidth for other users. It is also used in situations where local cell reception may be poor, allowing the user to connect via wired services with better connectivity.

The mobile offloading action can be set by either an end-user (mobile subscriber) or an operator. The code operating on the rules resides in an end-user device, in a server, or is divided between the two. End users do data offloading for data service cost control and the availability of higher bandwidth. The main complementary network technologies used for mobile data offloading are Wi-Fi, femtocell and Integrated Mobile Broadcast. It is predicted that mobile data offloading will become a new industry segment due to the surge of mobile data traffic

4. Rsearch Methodology

The information of user connectivity in the user mobility pattern is captured with the inter-contact model Specifically, as shown in Fig. 1b, for an arbitrary pair of mobile users, the timeline consists of both contact times, defined as the times when the users are within the transmission range, and inter contact times, defined as the times between contact times.



(a) A sample network.

(b) The timeline of user 1 and user 3.

Fig. 3. A sample network

Each file is encoded into several segments by the rateless Fountain code ,and a mobile user needs to collect enough encoded segments to recover the requested file. Once a user is in contact with some other users who cache part of its requested file, it will get some of the segments via D2D communication. Our objective is to design caching placement to maximize the data offloading ratio. The main contributions of this paper are summarized as follows:

- **1-**As each mobile user may get the requested file via multiple contacts with other users, the complexity of calculating the objective function, i.e.,the data offloading ratio,increases exponentially with the number of users,which brings a major difficulty for algorithm design.We first propose a divide and conquer algorithm to efficiently evaluate the objective,with quadratic complexity with respect to the number of users.

- **2-**The caching placement problem is shown to be NP-hard and a dynamic programming (DP) algorithm is proposed to obtain the optimal solution with much lower complexity compared to exhaustive search.Although the DP algorithm is still impractical, it can serve as a performance benchmark for systems with small to medium sizes. Moreover, its computation complexity is linear with the number of files, and thus it can easily deal with a large file library.

- **3-**To propose a practical solution, we reformulate the problem and prove that it is a monotone submodular maximization problem over a matroid constraint.The main contribution in this part is to prove that the complicated objective function is a monotone submodular function.With the reformulated form, a greedy algorithm is developed,which achieves at least 1 2 of the optimal value.

5. CONCLUSIONS

In mobility-aware caching, he exploited user mobility to improve caching placement in D2D networks using a coded cache protocol. He took advantage of the inter-contact pattern of user mobility when formulating the caching placement problem. To assist the evaluation of the complicated objective function, we proposed a divide and conquer algorithm. A DP algorithm was then developed to find the optimal caching placement, which is much more efficient than exhaustive search. By reformulating it as a monotone submodular maximization problem over a matroid constraint,he developed an effective greedy caching placement algorithm, which achieves a near-optimal performance.

6. REFERENCES

1. X. Peng, J. Zhang, S.H. Song, and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in Proc. IEEE Int. Conf. Commun. (ICC), Kuala Lumpur, Malaysia, May 2016.

2. R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: modelling and methodology," IEEE Commun. Mag., vol. 54, no. 8, pp. 77 – 83, Aug. 2016.

3. Y. Shi, J. Zhang, K. B. Letaief, B. Bai, and W. Chen, "Large-scale convex optimization for ultra-dense Cloud-RAN," IEEE Wireless Commun., vol. 22, no. 3, pp. 84–91, Jan. 2015. V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell

4. R. Lan, W. Wang, A. Huang, and H. Shan, "Device-to-device offloading with proactive caching in mobile cellular networks," in Proc. IEEE Global Commun. Conf. (GLOBECOM), San Diego, CA, Dec. 2015.

5. B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in Proc. IEEE Int. Conf. on Commun. (ICC), London, UK, Jun. 2015.

6. H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," IEEE/ACM Trans. Netw., vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

7. Passarella and M. Conti, "Characterising aggregate inter-contact times in heterogeneous opportunistic networks," in Proc. IFIP Netw. Conf., Valencia, Spain, 2011.

8. G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák, "Maximizing a monotone submodular function subject to a matroid constraint," SIAM J. Comput., vol. 40, no. 6, pp. 1740–1766, 2011.

9. S. El Rouayheb, A. Sprintson, and C. Georghiades, "On the index coding problem and its relation to network coding and matroid theory," Information Theory, IEEE Transactions on, vol. 56, no. 7, pp. 3187–3195, Jul. 2010.

10. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," IEEE Trans. Mobile Comput., vol. 6, no. 6, pp. 606–620, Jun. 2007.