New Collaborative Caching Scheme for D2D Content Sharing in 5G

Rana E. Ahmed
Dept. of Computer Science and Engineering,
American University of Sharjah, United Arab Emirates
Email: rahmed@aus.edu

***Abstract*—**Device-to-Device (D2D) content sharing is an excellent solution to offload the traffic from the core 5G networks and backhaul links. Caching the most popular files at the network edge and in users' devices to support user proximity services in 5G helps to offload traffic in the core network and increases the cache hit probability. D2D communication in 5G can be utilized to share the cached files between a pair of devices with a minimal involvement from the base station. This paper presents a collaborative content caching scheme with the primary objective to maximize the overall system offloading gain and the cache hit probability in downloading the popular file contents. The proposed system model exploits the social-networking concept within the WiFi Direct radio range, assuming the cell structure present in a densely populated area such as a university campus or an auditorium. The performance of the proposed cache scheme is evaluated through extensive simulations taking into consideration the device mobility and disconnection, and the results are presented.

***Index Terms*—**D2D communications, 5G, content sharing, caching

I. Introduction

Device-to-device (D2D) communication allows users to exchange information directly with each other in an ad hoc manner with no or very limited involvement of the Base Station (BS). D2D communication offers several advantages, such as higher power efficiency, spectrum efficiency, and lower delays. Another important advantage of D2D communication is in the caching strategies in 5G where contents can be stored at the edge of the network that can be later shared among several User Equipment (UEs) using D2D. This strategy of contents sharing also reduces traffic tremendously on the core network and backhaul links.

Caching the most popular files at the network edge is a promising technique that will be utilized heavily in the era of 5G. Caching is used to reduce the need for accessing the data directly from the origin server at the content provider end, which will reduce the core network congestion and hence reduce the overall latency, while increasing the transmission data rate. In 5G, caching will help in reducing the network load by storing the most popular files at the network nodes such as the Evolved Packet Core (EPC), Radio Access Network (RAN), Base Station (BS) and the UEs. Caching can be done at the off-peak hours to reduce the overall latency. Caching at the UEs exploits the D2D links (e.g., WiFi Direct, Bluetooth) among various user terminals. With today's advances in memory technologies, a UE device can have reasonably large storage capacity; therefore, by using the D2D communication link, these devices can share the popular content with their peers. The key factor of success in this operation is to pick the perfect set of files based on their expected popularities so that the request of users can be satisfied directly at the network edge without the need of using the congested backhaul link to the Content Provider [1].

Four different types of file caching and offloading mechanisms used are shown in Fig. 1. The self-offloading mode represents the scenario when a certain user requests a file that is already available in its storage. In such case, the requested contents can be offloaded directly and the request is satisfied immediately. In the Macro Base Station (MBS) offloading mode, the requested contents are not found within the device itself or any of the neighboring devices. The contents are then supplied by the origin contents provider/ server via the MBS. In the D2D offloading mode, a user requests a file that is available at any neighboring device; in this case, the request can be satisfied by utilizing the available D2D communication between the two devices. For any such file request there may be more than one potential providers; in such case, the requester can make multiple D2D links with the multiple providers to satisfy the request. In the Small Base Station (SBS) offloading mode, the network is a HetNet with two-tier network that consists of a Small Base Station (SBS) and the Macro Base Station (MBS). In such case, the SBS forms a cache helper, and the user request can be directly satisfied by the SBS [1]-[5].

The authors in [6] have divided the macrocell into several virtual clusters, and if the content file is not available within the cluster, the MBS will satisfy the request by getting the file from any neighboring cluster and then deliver the file. Both the centralized and distributed cache schemes within a certain collaborative distance are examined. It is concluded that the random cache placement, where the location of the devices is unknown a priori, shows only minimal loss in the performance.

Manuscript received December 16, 2021; revised June 15, 2022.
doi:10.12720/jcm.17.6.521-527

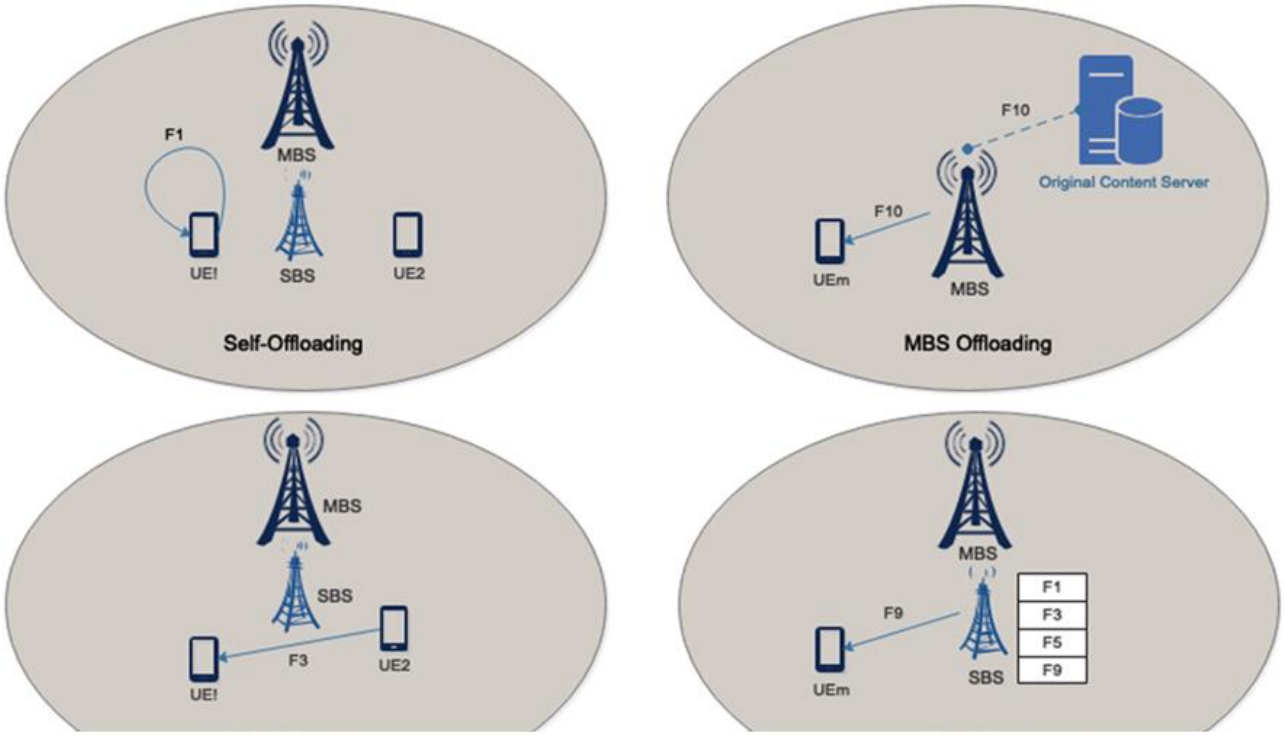

Fig. 1. Different types of off-loading requests in 5G networks (adapted from [2]).

Two popular performance criteria used in the cache-enabled D2D communication schemes are: cache hit probability and the offloading probability. The cache hit probability is the fraction of the total system requests for contents satisfied by one or more UEs caches. Higher cache hit probability indicates that most of the requests are successfully serviced locally, and the UE needs to connect to the BS a few times.

The offloading probability (also known as the offloading ratio) is the ratio between D2D traffic and overall traffic (D2D + Cellular) for transmitting all the messages requested.

This paper proposes a new collaborative caching scheme for D2D communications. The scheme issues the environment of social networks where the end users are in a close proximity, such as office, home, or university campus. The scheme uses the WiFi direct for the D2D communication, and it uses new content placement and replacement policies based upon the content popularity.

The rest of the paper is organized as follows. Section II presents a brief overview of the problem and briefly describes the major issues in D2D content caching. Section III describes the system model used for the proposed scheme. Section IV describes the proposed scheme in detail. Section V discusses the simulation environment and the results for the proposed scheme, while conclusions are drawn in Section VI.

II. Issues in D2D Content Caching

We assume that the multimedia content of interest is composed of one or more files. The file size can be determined by considering the storage capacity available at a device or network element. There are several issues that are needed to be addressed for any caching scheme to be effective. The issues can be broadly categorized in the following policies: Cache placement. Cache content access, Cache replacement, and Error handling policies.

A. Cache Placement Policy

This policy attempts to find solutions for the following important questions:

- Where are we going to place the multimedia contents? Will it be centralized (i.e., placed at only one location) or distributed (i.e., placed at multiple locations in a collaborative way)?
- Which UE(s) or network elements will hold the contents?
- What events will trigger the placement of contents at a specific network element?

B. Cache Content Access Policy

This policy tries to find solutions for the following questions:

- How will the cache contents be accessed in a cost-effective manner? The cost could be a function of delay, channel bandwidth, among many others.
- What criteria a device needs to use to maximize the cache hit probability?
- How will the popularity index for a certain file change?

C. *Cache Replacement Policy*

Since the cache memory area for content storage at a device is limited, one of the existing files in the device cache will need to be deleted (purged) in order to make room for a newly referenced file. This policy decides which existing file needs to be replaced. The replacement criterion is based on the popularity index of the file. A file can be replaced if it is not popular anymore, and/ or

the probability of accessing this particular file will be quite low in the future.

D. Error Handling Policy

In order to deal with the non-cooperative behaviors from nodes when they unilaterally delete the contents or do not cooperate in the file access, there should be a well-defined error handling policy present as a part of content caching strategy. This non-cooperative behavior could be due to some legitimate reasons, e.g., low residual battery, poor coverage, etc. The non-cooperative behavior is modeled as average disconnection rate in this paper.

It should be noted that some features of the above-mentioned policies may have conflicting requirements. The main overall goal of the proposed scheme is to maximize the cache hit probability and minimize the file transfer delay.

III. System Model

It is assumed that the D2D communication does not interfere with the communication between the base station (BS) and other users. This assumption is justified if the D2D communications occurs in the unlicensed frequency band (e.g., WiFi). We assume that groups of mobile devices collaborate to exchange files via D2D communications. We can say that clusters of collaborating devices "pool" their caching resources to provide a "central virtual cache", controlled by the BS, and a user can select from multiple devices to form a D2D pair.

When a UE first joins the network and requests for a content file, it directs its request to the BS, which keeps a record that contains the cached files IDs and the related owners of those files. The BS sends the requester the information about the possible content owners, and then the requester attempts to connect to one of the possible owners that falls in the connectivity range and the communication channel has the highest SINR value. The proposed model is a hybrid content caching assisted D2D communication, since the BS is partially involved in the process. The model takes into account the user's mobility in addition to the random nature of users' behavior for request arrivals, departures, and disconnection.

A. Network Model

We consider a 5G network cell that consists of the MBS and a set of UEs which are distributed under the coverage of the MBS following the independent homogeneous Poisson point process (PPPs) with density of λ in the unit of number of users per unit area (users/m^2). A pair of UEs will discover each other and establish a D2D communication channel in the downlink to exchange the cached contents. The distance between the D2D transmitter and receiver is limited to 50 m, as stated in the 3GPP standard Release 12 [7]. We assume that the D2D is operating in the "unlicensed spectrum", and using the WiFi interface, as is done in [8]. We assume a D2D peer discovery is done using LTE Direct technology available at many smart phones today. The MBS is equipped with only one antenna for transmission and reception and the same is assumed for a UE. We assume that each D2D pair can operate either in the D2D mode or in the cellular mode. In our work we focus mainly only on the D2D mode of operation. The selection of one of the modes or another depends on the specified criterion. The most important criteria to be met is the availability of the requested content at the transmitter's cache memory. If the content is available, then a possible D2D pair can be established between the transmitter and the receiver of the content. In cases when the requested content is unavailable at a neighbouring UE, then the request will be directed to the BS and the mode will be switched to the cellular mode. We assume that the network covers a dense auditorium or university campus, the interference is high, and the network is congested with the download requests.

B. Contents Cache Model

For the cache placement problem, we assume a set of content consisting of N number of files. For simplicity we assume all files are of the same size. The whole file set is also cached at the BS, portion of the set is distributed to be cached at some of the UEs. The UEs has a limited cache storage capacity, which means that not all the files can be cached in a UE. Furthermore, the BS keeps track of which files are cached at which UE device using a binary matrix. There are three modes of contents transfer: Self-offloading, D2D and B2D (Base Station to Device).

In the Self-offloading mode, the requested content file is already available at the requesting device. In this case, there will be no need to connect with any neighboring device and the request will be served immediately. This is the case where a user is revisiting a previously requested content.

In the D2D mode, the requested content file is available at one or more of the neighboring devices, which form the potential content providers, and the connection can be established to connect with one of the content providers. The requester then sets up a one-to-one connection between itself and the content provider, and the request is then satisfied.

In the B2D mode, the requested content file is not available at any of the neighboring device, and it can be only available at the BS. In this case, the requesting device utilizes the cellular link to establish a one-to-one connection between itself and the BS. If the requested content is also not available at the BS, which is the worst-case scenario, then the BS utilizes the backhaul link to bring the requested content from the original content service provider. In this paper, we assume that the requested content has been already fetched from the remote server.

IV. The Proposed Caching Scheme

We present the proposed algorithms for a distributed and decentralized content sharing between UEs by utilizing the available D2D links. The major steps for the proposed algorithm are:

- D2D pair selection

- Cache-Assisted D2D Communication
- Cache contents placement and replacement strategies
- Handling disconnections and other anomalies.

For the D2D pair selection, the proposed algorithm uses a strategy based on the Maximum Signal-to-Interference Ratio (SINR) and the minimum distance to select the D2D pair to communicate with. If one or more devices are within the communication range of the requesting device, then the algorithm will select the one with the higher SINR on the communication channel, regardless of the distance between the two devices, as long as the distance allows for the D2D communications. The SINR is an indicator of the channel quality (CQI) and it is calculated within a UE [9].

For the cache-assisted D2D communication phase, based on the (un)availability of the requested content file cached at one or more neighboring devices, a potential mode of operation will be selected: i.e., self-offloading, D2D mode or B2D mode. The B2D will be selected only if the content could not be found in any neighboring user and/ or the communication between the requesting device and the potential provider cannot be established.

On the occurrence of the first request for a new content file, the request is placed by the user to the BS. The BS serves the request using the backbone network, and the file is cached at the BS. The file is then transmitted to the file requester, and then the requester caches the file as well. Based on how many times the file is requested by other users, the file popularity index is calculated locally at the content owner. The replacement of the file in the existing cache storage is done primarily using the Least Frequently Used (LFU) algorithm. After sorting the files in a stack of descending order based on the popularity index, the last file in this stack will be replaced.

The content popularity distribution normally follows a Zipf distribution, which is widely used to model the multimedia content popularity [10]. The proposed scheme assumes that the files popularity follows the Zipf distribution with skewness of η. We consider a content storage consisting of *N* files that all users may request. The files are indexed in descending order of their popularity, i.e., the first file is the most popular one.

The probability, P_k , that *k*th file is requested follows the Zipf distribution with parameter Η, is given by

$$P_k = \frac{k^{-\eta}}{\sum_{l=1}^{N} l^{-\eta}} \quad (1)$$

and where

$$\sum_{k=1}^{N} P_k = 1 \quad (2)$$

The parameter Η reflects how skewed the popularity distribution is, with large Η meaning that a few files are responsible for the majority of requests.

The replacement method depends not only on the Least Frequently Used (LFU) strategy, but also on the number of owners to that least frequently used (i.e., potential victim) file. If the LFU file is cached only at one UE, then it will not be replaced, and the algorithm will keep on looking into a file that has a lower popularity that is owned by more than one UEs. This process tries its best to keep a file that has only one owner in the system. On the other hand, a file which resides in more than one user cache storage can be replaced since it has more than one replica. This replacement strategy has the potential to increase the overall system efficiency, and hence will increase the overall system caching gain.

In order to provide a realistic model that simulates the real-life events, the proposed scheme takes into account the mobility of the users. The unpredictable users' mobility affects the cache hit probability and the cache offloading mode, i.e. D2D mode and B2D mode. The proposed scheme uses the Random Walk Model (RWM), which is a popular model in mobile wireless communications [11]-[14].

V. Simulation Results and Discusion

The proposed caching scheme was simulated under MATLAB environment. Table I shows the major system parameters with their default values used in the simulation.

Two performance metrics investigated during simulation are: cache hit probability, and offloading probability. Cache hit probability is the fraction of file requests that are satisfied by the device caches, either through self-offloading or D2D mode. The offloading (or, more specifically, D2D offloading) probability is the fraction of file requests that are satisfied through D2D communication only, without going through the communication to the BS.

TABLE I: MAJOR SYSTEM PARAMETERS WITH THEIR DEFAULT VALUES

Parameter	Default Value
Cell radius	1500 m
Channel Model	Macro Urban
D2D Distance Threshold	50 m
Density of users	5×10^{-5} /m^2
Request rate for the system	0.5 request/s
UE Transmit power	+24 dBm
Content library size	100 files
Cache storage size	10 files
Size of content	1 Mbits
Carrier frequency for D2D	2.4 GHz
Bandwidth for D2D	22 MHz
Path loss exponent (α)	4
Zipf distribution parameter	2
SINR threshold	15 dB
UE Receiver sensitivity	-100 dBm
AWGN noise	-117 dBm

Fig. 2 compares the cache hit probability versus different values of the D2D distance threshold. D2D distance threshold is the maximum allowable cooperative distance between any two D2D devices. Form the figure it can be clearly seen that the hit probability increases with the increase in distance and then saturates after about 50m.

This is because the larger communication distance will allow more devices to communicate (due to larger coverage area), and hence will introduce additional interference to the system. Furthermore, the increase in the communication distance will affect the channel quality which will result in many connection failures. On the other hand, for small values of D2D distances, the increase on the Zipf parameter η increase the hit probability, since an increase in the Zipf parameter will increase the probability that the file is cached in one of the devices. As a result, the requester will have a higher chance in getting the requested file from nearby devices.

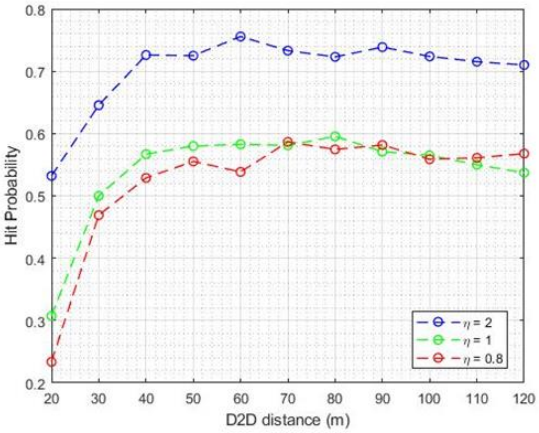

Fig. 2. Cache hit probability vs. D2D distance with different Zipf parameter η.

Fig. 3 compares the hit probability against different values of the Zipf parameter, while varying the cache size. The hit probability increases as the cache storage size increases. This is due to the fact that increasing the Zipf parameter will increase the cache decision probability. Moreover, the large cache capacity will improve the chances that a requested file is cached in the cache storage, which in turn increases the cache hit probability. As a result, more requests will be offloaded from the D2D link rather than from the BS.

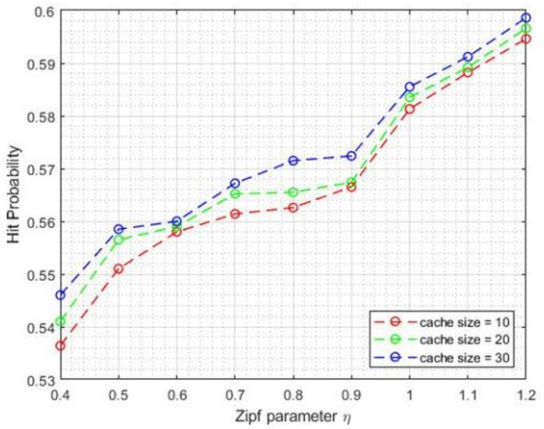

Fig. 3. Cache hit probability vs. Zipf parameter η with different cache sizes.

One of the contributions of this research work is to study the effect of user's unpredicted disconnection from the communication cell. Fig. 4 shows the values of the cache hit probability without any disconnection, with low disconnection rate of 10%, with average disconnection rate of 50% and with a very high disconnection rate of 90%. The effect of user's disconnection on the hit probability is severe in the average and high values of disconnection rates, as a result there could be a need of a mechanism to minimize the degradation is performance due to the user disconnection. One of the proposed solutions could be some sort of incentive-based mechanism to encourage users to continue participating in the cache-assisted D2D communication.

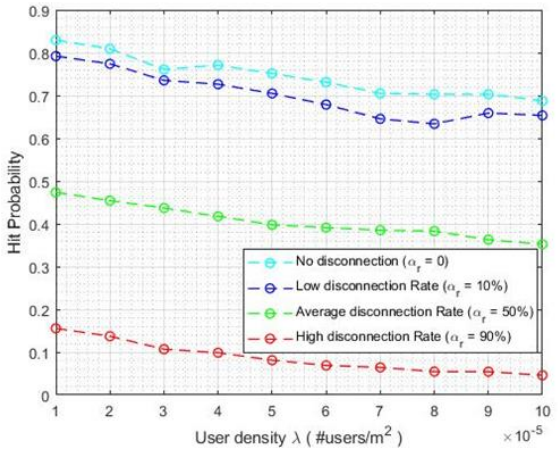

Fig. 4. The impact of user's disconnection as under different values of the user density λ (users / m^2).

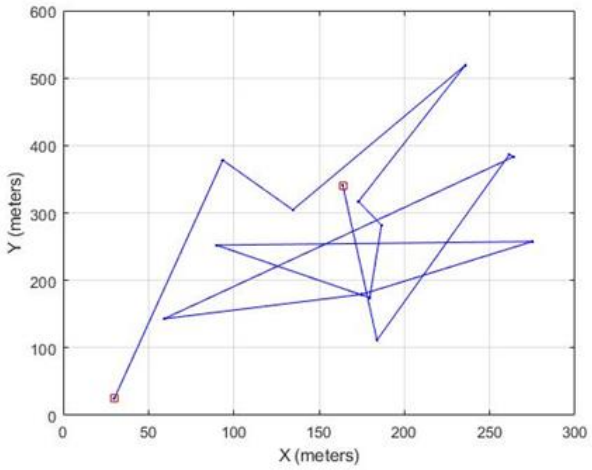

Fig. 5. Sample node movement in RWM model.

The effect of device mobility to overall system offloading gain has been also investigated. A sample device movement using the Random Walk Model (RWM) is shown in Fig. 5. The effect of user's mobility in the cache offloading probability is shown in Fig. 6. It has been found that the effect of mobility in the offloading gain is minimal in small speeds till 10 km/h. The percentage of requests offloaded using D2D connection

increases as the speed increases. At a low speed of 3 km/h the offloading probability was found to be 60%, and this percentage increases to 89% when the speed reaches 40 km/h. The low speed of 3 km/h simulates the average speed of a pedestrian, while the moderate speed of 40 km/h simulates the average speed of a user riding a motor bicycle. The increase of user's speed makes the user appears in several different locations, allowing better chances for other users to communicate and access the content.

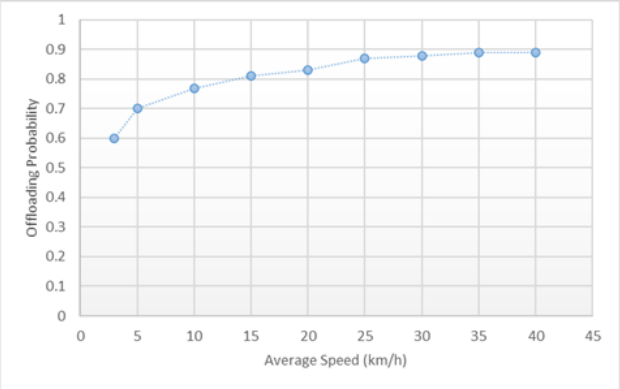

Fig. 6. The impact of mobility on the offloading probability.

VI. Conclusion

Caching at the edge of the network with the assistance of D2D communication is presented as an emerging strategy in 5G networks to cope with the ever-increasing demand in multimedia services. The main goal behind this technology is to increase the cache hit probability which in turn increases the overall offloading gain.

This paper presents a collaborative content caching scheme with the primary objective to maximize the overall system offloading gain and the cache hit probability in downloading the popular file contents. The proposed scheme exploits the social-networking concept within the WiFi Direct radio range, assuming the cell structure present in a densely populated area, such as a university campus or an auditorium. The performance of the proposed cache scheme is evaluated through extensive simulations. The effect of user density in the cell, disconnection rate and the mobility are also presented. It has been found that the proposed scheme works well in normal scenarios, and effects of low-to-moderate device disconnection and mobility do not drastically impact the cache hit probability and offload gain.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgement

The author acknowledges the contribution of his student, Ms Ansam Abdelsalam, for conducting some simulations.

References

[1] D. Prerna, R. Tekchandani, and N. Kumar, "Device-to-device content caching techniques in 5G: A taxonomy, solutions, and challenges," *Computer Communications*, vol. 153, March 2020, pp. 48-84.

[2] I. Parvez, A. Rahmati, I. Guvenc, A. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, Core network and caching solutions," arXiv: 1708.02562v2, May 2018.

[3] D. Wu, L. Zhou, Y. Cai, and Y. Qian, "Collaborative caching and matching for D2D content sharing," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 43-49, June 2018.

[4] S. Hosny, A. Eryilmaz, A. Abouzeid, and H. El Gamal, "Mobility-aware centralized D2D caching networks," in *Proc. 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Feb. 2016.

[5] S. Raza Zaidi, M. Ghogho, and D. McLernon, "Information centric modeling for two-tier cache enabled cellular networks," in *Proc. IEEE International Conference on Communication (ICC) Workshop*, London, UK, June 2015.

[6] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis, "Base-Station assisted device-to-device communication for high-throuput wireless video networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, July 2014.

[7] R. Wang, J. Zhang, S. Song, and K. Letaief, "Mobility-Aware caching in D2D networks," *IEEE Transactions on Wireless Communications*, August 2017.

[8] S. Iskounen, T. Nguyen, S. Monnet, and L. Hamidouche, "Device to Device communications using WiFi Direct for dense wireless networks," in *Proc. IEEE 7th International Conference on the Network of the Future (NOF)*, Jan. 2017.

[9] F. Afroz, R. Subramanian, R. Heidary, K. Sandrasegaran, and S. Ahmed, "SINR, RSRP, RSSI and RSRQ measurements in long term revolution netwroks," *International Journal of Wireless and Mobile Networks*, vol. 7, no. 4, Aug. 2015.

[10] D. Malak and M. Al-Shalash, "Optimal caching for device-to-device content distribution in 5G networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 863-868.

[11] M. Chen, Y. Hao, L. Hu, K. Huang, and V. Lau, "Green and mobility-aware caching in 5G networks," *IEEE Transactions on Wireless Communications*, Dec. 2017.

[12] V. K. Quy, L. N. Hung, and N. D. Han, "CEPRM: A cloud-assisted energy-saving and performance-improving routing mechanism for MANETs," *Journal of Communications*, vol. 14, no. 12, pp. 1211-1217, 2019.

[13] R. E. Ahmed, "A low-overhead multi-hop routing protocol for D2D communications in 5G," *Journal of Communications*, vol. 16, no. 5, pp. 191-197, April 2021.

[14] V. K. Quy, N. T. Ban, V. H. Nam, D. M. Tuan, and N. D. Han, "Survey of recent routing metrics and protocols for mobile ad-hoc networks," *Journal of Communications*, vol. 14, no. 2, pp. 110-120, 2019.

Rana E. Ahmed received his PhD in Electrical Engineering from Duke University, USA, in 1991. He is currently working as a Professor at the Dept. of Computer Science Engineering, American University of Sharjah, Sharjah, United Arab Emirates. He has worked as a faculty member at Lakehead University, Canada, and at King Saud University, Saudi Arabia, in the past. He was a visiting faculty member at the University of Ottawa during academic year 2007-08 while on sabbatical leave. He also worked at Research in Motion (RIM) (now known as BlackBerry) and SpaceBridge in Canada in the areas of software testing, and software quality assurance. His research interests are in the areas of computer networking, computer architectures, fault-tolerant computing, and software engineering.