# USE OF CACHING IN HETEROGENEOUS CELLULAR NETWORK

**[1]Shroog Ageel Alshammari, [2]Reem Abdullah Alkhateeb, [3] Kusum Yadav**

*Computer Science and Engineering Department, University of Ha'il.* Hail, Kingdom of Saudi Arabia

[1]shr.alshammari@uoh.edu.sa, [2]r.alkhateeb@uoh.edu.sa, [3]kusumasyadav0@gmail.com

***ABSTRACT—*** *Caching has appeared as an important technique in modern communication systems for defeating peak data rates by providing popular files to be pre-fetched and cached locally at end-users' machines. With the shift in the model from homogeneous networks to the heterogeneous ones, the idea of data offloading to small cell base stations (SBSs) has earned important observation to avoid bottlenecks in the moderate capacity backhaul link to the core network. In this paper, we implement a multicast-aware caching approach for small-cell networks in such a heterogeneous small cell wireless network to minimize the energy costs. This multicast-aware caching problem is an NP-hard problem. However, their heuristic approach has been implemented to overcome the hardness issue and achieve significantly lower traffic.*

## I. INTRODUCTION

The recent generation of smartphones has considerably enhanced the mobile user experience, leading to a broad range of new wireless services, including web-browsing applications, multimedia streaming, and socially-connected networks. This aspect has caused the mobile data traffic to grow dramatically over the last decade, and since the number of connected devices keeps increasing, there will be 11.6 billion mobile-connected devices by 2021 which will exceed the world's proposed population at that time (7.8 billion according to Cisco statistics [1]). Simultaneously, social networking is already the second-largest traffic size contributor with a 15% average share. Consequently, mobile data traffic is expected to continue growing to reach 49 Exabytes per month by 2021 [1]. This essentially means a sevenfold increase over 2016 and the mobile data traffic will grow at Compound Annual Growth Rate (CAGR) of 47 percent from 2016 to 2021. Ensuring the reliability and scalability of modern-day networks to manage such massive amounts of data is an active research area. In the meantime, standard wireless networks have reached their capacity limits, particularly during peak hours.

This new phenomenon has made mobile operators to redesign their current networks and investigate more advanced techniques to enhance coverage, increase network capacity, and cost-effectively fetch contents closer to users. This difficulty has been faced by increasing the data throughput via more radio spectrum plus adding multi-antenna techniques, and performing more suitable modulation. However, these criteria alone are not sufficient to keep up with the most crowded environments where performance can significantly corrupt [2].

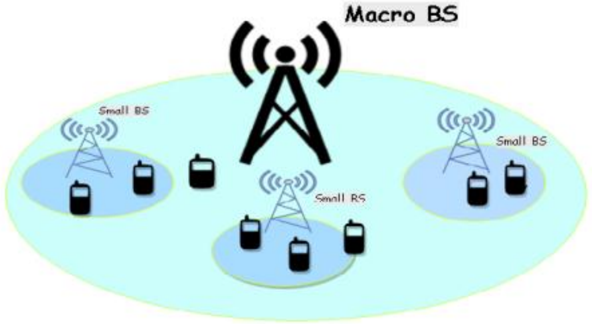

**Fig.1. A typical Heterogeneous Network (HetNet).**

A proper approach to meet these unprecedented traffic demands is via the deployment of Heterogeneous Cellular Networks (HetNet). The HetNet is a recent mobile communications network composed of a combination of various cell and various access technologies. Typical HetNet, as illustrated in Figure 1, consists of conventional Base Stations (BSs) and Small cell Base Stations (SBSs) (i.e., micro-, pico-, and femtocells).

HetNet represents a novel networking model based on the idea of deploying low-power, short-range, and low-cost SBSs. HetNet based on SCs has recently been developed to fulfill the huge future demands and enhance the standard network capacity and the Quality of Experience (QoE) [2]. The primary objectives of a HetNet are: reaching larger coverage and higher capacity while contributing seamless handoff and connectivity. SCs have been proposed to add low-cost network capacity through aggressive reuse of the cellular spectrum. The SBSs implement a cost-effective way to offload the traffic from the central macro-cellular networks [3]. The use of SBSs overcomes the need of all the communication to cross through conventional BSs. Nevertheless, the possible performance that is assumed from the deployment of the SBSs will be limited by capacity and heterogeneous backhaul links which connect the SBSs to the core network [3]. The main bottleneck is expensive of the backhaul link that connects the macro base station with the SBSs.

To deal with the backhaul bottleneck, the content providers are moving their users' content to the intermediate nodes in the network, namely caching, producing fewer delays for the access. Distributed caching at the network edge is considered as a promising solution. The basic concept of caching is to duplicate and store the files at the SBSs side to serve the clients. Thus, without using the backhaul links, all users' demands can be served from the nearest SBSs as much as possible.

In this paper, a general introduction and background to the SCs and HetNet are provided in section two. Section three presents a general concept of Caching. Next, in section four, we discuss related works about caching contents. Then, we model the cache problem in SBSs as a multicast caching problem, and we implement a heuristic algorithm that aims to reduce the backhaul load. Following that, the section shows the results obtained from the algorithm. Finally, the conclusion and future work are discussed in section seven.

## II. GENERAL OVERVIEW ON SBSS

Due to the proliferation of smartphones and tablets, the traffic load in the current cellular networks is showing no signs of slowing down. Most of the traffic load is caused by multimedia content that demands a high Quality of Service (QoS). To keep bearable delays and to efficiently utilize and to meet the future user equipment (UE) traffic demands, novel strategies for modeling wireless networks are needed. According to [4], the traditional networks have met their capacity limits. As the earlier used solutions to improve the capacity of the network has already met the ergodic capacity limit, many smaller cell sites are required. It is now well known that large BS provides coverage for a large area and consumes a large amount of power but the effective way to improve network capacity is making cells smaller by reducing the path to the users. In particular, the small cells include all low-powered radio access base stations which work in a range spectrum of 10 meters to many hundred meters in contradiction to the macrocells which work in a range up to many tens of kilometers. They are small compared to macrocells, partially because they have a shorter range and partially because they manage fewer concurrent calls or sessions. Several various terms have been used across various parts of the industry,

From the sixties, the idea of caching was known in the context of algorithm design in operating systems [7]. With the discovery of content delivery networks (CDNs) and the exploitation of web caching, the bottleneck eased. The optimal content excluding strategy in the case of new content coming is to remove that content from the memory which is not going to be demanded shortly. In the past decades, there have been studies on web caching methodology trying to improve the scalability of web and offloading the network. Web caching replicates popular content in various geographical spaces driving to bandwidth savings by avoiding unnecessary multi-hop retransmissions. This also reduces access time/ latency by decreasing the distance for accessing demanded content [8]. Additionally, the information-centric networking (ICN) which is a candidate for the Internet of the future, intends to alter the method of accessing the data on the internet. ICN asks the essential question of where to place network memory for greater benefits. To this end, ICN moved to prepare routers with caches to smooth network-wide content replication. It is noted in [9] that maximum caching gains in ICN could be achieved by caching at the edges of the network using present CDNs.

Some studies have recorded that a global caching profit can be invested in wireless networks. As mentioned before the network caching consists of the delivery phase, which is a stage that satisfies the demands made by the receivers during the repeat of the network. The delivery phase of cache-aided systems performed as a set of dedicated point to point unicast transmissions to particular users by delivering portions of inquired files which are not stored in their storage yet is not a scalable answer since the number of users in the system is raising. Some of the previous works in this region tend to adopt a fixed transmission system after that; they optimize the storage phase to accommodate the delivery scheme [10, 11]. There, researchers are ignoring the scope of global cache interactions across users and content and primarily based on the gains achieved by local content distribution. These recent works [12, 13] by Ali and Niesen, included theoretic formulation information about caching in wireless networks. It is exhibited in these works that by jointly composing the storage with the delivery phase by applying multicast transmissions to deliver content to the receivers simultaneously, improvement in the transmission rates can be obtained as compared to traditional unicast delivery. Based on the shared content in user caches, the caching and multicast delivery scheme was introduced, whereby a global caching gain obtained from the system adds to the traditional local additions. To derive lower bounds on optimal cache storage vs. transmission rate trade-off, the authors used the cut-set based arguments and defined it to within a constant multiplicative factor of 12. In the case of video distribution, the proposed algorithms are elegant due to their relative smoothness of implementation and efficiently rekindled the importance in cache aided networking from an information-theoretic viewpoint.

Recently, it has been noticed that a substantial part of mobile data traffic is caused by multiple duplicate downloads of some popular content [21]. However, the storage capacity of today's memory devices has increased quickly at a comparatively low cost. Motivated by these facts, Golrezaei *et a*l. [15] proposed to replace SBSs with the BSs that have limited backhaul links or without backhaul but have high capacity caches, called helper nodes. They provide femtocells with large caching capacity and combine to the network, helper nodes. By optimizing the caching policies and storing the most popular files to serve more users under the constraints of file downloading time, in the case of unavailability of data in their cache, the MBSs serve the users. As a result, significant throughput gain was recorded. Furthermore, regarding small cell networks (SCNs) with backhaul of insufficient capacity, the authors in [16] noticed that the backhaul traffic load could be overcome by caching files at the SBSs based on their popularity. It indicated that by getting content locally instead of taking from the core network through backhaul links redundantly, implementing caches at BSs is a promising approach to unleash the potential of HetNet. Nevertheless, the performance of a cache-enabled HetNet beyond a standard HetNet with limited capacity backhaul is still unexplored [17].

Caching as an aide to wireless interference mitigation was investigated [18]. In [19], it is shown that the gains from caching and massive MIMO are indeed complementary. The authors in [20], derived caching at relays, the throughput of a cache-enabled network with content shifting to users and D2D communication, where all the nodes including MBS and relay are with high capacity backhaul and with an individual antenna. Chae *et al*. [21] analyzed the probability success of the file delivery of cooperative transmission amongst helper nodes, where these helpers and BSs are spread in orthogonal bandwidth.

To reduce service delays, [22] suggested a content caching scheme and shared user clustering as a method of accounting for the difference in content popularity among different users. This approach utilizes the common similarities among users, regarding users' attention in contents. The similarity matrix is built and then used to group users and assign them to small cells which are provided with cache memory, within a training period. In the [24], a preliminary study on transfer learning for caching in SCNs is conducted. Bastu˘g, *et al*. show that the content popularity estimation through CF can be enhanced by this approach. The author in [25] utilized the centrality measures for the content placed in the context of proactive caching. Users tend to esteem highly suggested content by people who have similar interests. Thus Twitter, Facebook, and some other online social networks have become useful in distributing several contents over social communities [26].

Consequently, it can alleviate peak traffic demands by benefitting from users' social relationships, and proactively storing the content on users' devices. The authors in [23] studied the content popularity estimation depending on the aggregated context information such as user density, the day time and also depending on the user characteristics, such as gender or age. In contrast, [27] presented the context information of an individual user rather than aggregated context information. Moreover, this approach studies more than the diversity of content popularity; with different users, it reacts to arbitrary user entries concerning their contexts. Lots of other related research suppose a priori knowledge of the popularity of the content. In most cases, the contents' popularity profile is illustrated by a Zipf distribution [26]. The traditional caching policies place the most popular contents in the cache memory. Thus they are popularity-aware. The wireless medium can be exploited to enhance the caching gain through coded caching [28]. The multicast transmission of identical contents which are requested in a short time window allows energy saving. This achieved in [29] by proactive caching. Further, Hachem et al. [30] exploited the broadcast capability of MBS and the storage capacity of SCs in dense wireless networks and introduced a good trade-off within the transmission cost of the BSs, the cost of connecting users to multiple SCs and the storage cost of the SCs.

## III. CHACHING POPULAR FILES

The success and popularity of YouTube, HBO, Netflix, and other video content providers have contributed to the growth of network traffic. To solve this, content providers have sets files stored in their data servers, and they aim to cache the files closer to the users in SBSs, to reduce the delivery time without lags or interruptions. However, the limited backhaul capacity is recognized as one of the most significant issues in the SBSs. Thus, the providers intend to cache the most popular files to overcome the backhaul load in the network.

It is proposed to start by caching the most popular files. For example, the files that have the highest probability to be requested by the users and depending on the preferences of users that are connected to each SBS. Over time, the popularity of files differs from one SBS to another SBS. However, to prepare files that should be cached in each SBS to reduce the cost of serving the requests is the main challenge. We study multicast caching transmissions to serve multiple requests of different users' strategies at the same time in the heterogeneous network. The purpose is to reduce the servicing cost of the operator by decreasing the volume of the incurred traffic. Multicast is the phrase used to define communication where a part of the information is transmitted from one or more points to a collection of other points. In this situation, there may be one or more senders (MBS, SBSs), and the information is assigned to a set of receivers (there may be no receivers or any number of receivers). As an example, an application that may use multicast is a video server transmitting out networked TV channels.

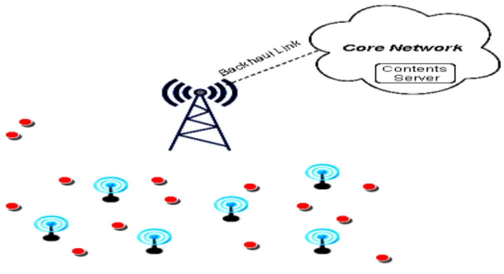

**Fig. 2. System Model**

The caching problem changes when multicast is applied to serve simultaneous requests for the same content file. In contrast to unicast communication, the multicast provokes less traffic as the demanded file is sent to users only once, rather than with multiple point-to-point transmissions. Where there is a common demand for the same data requested by a group of clients, the multicast transmission may provide vital bandwidth savings compared to N separate unicast clients.

### *A. **Assumptions***

We consider a network in which operators employ multicast transmissions to serve multiple requests of several users concurrently. This multicast delivers content to users who submit a file request at a nearby time. In this model, let's consider 6 of small cell base station $\mathcal{n}$, and these SBSs are connected to one macro base station MBS. Serving request of 100 users, U = {U1, U2,. . . U100}. Those users are located in an area covered by SBS or by MBS. Each SBSs $\mathcal{n} \in \mathcal{N}$ is provided with a cache size $\mathcal{S}n \geq 0$ . Please see Fig. 2

When the data is transferred from the core network to the MBS via the backhaul link, the energy cost will be C$b \geq 0$. Moreover, when the data is transferred from the MBS to the users, the cost is C$w \geq 0$. The energy cost is C$n \geq 0$ when the data is transferred to the closer user from the SBSs. We set C$b$ = C$w$ = 0.5 and C$n$ = 0. In our scenario, we assumed that all files $I$ have the same size and $I$ = 1000. Also, λni ≥ 0 when the users are in the coverage area with the average demand of users for file $\imath$ covered by SBS $\mathcal{n}$. However, λ0i ≥ 0 when the average demand of the users for file $\imath$ is not in the coverage area of any of the SBSs. In our model we set λ0i = 0. Further, the time deadline $d$ =20 seconds and $\mathcal{S}n$ = 24 units. The probability that at least one demand for file $\imath$ is generated by the users located in the coverage area of SSBS,it is denoted by $Pni$ in [1]. To this end, the popularity distribution of the files follows the Zipf law, $\alpha$ = 0.8.

### *B. **Formulation***

We introduce the $xni$ as the binary decision. Consequently, $xni$=1 if the file $\imath$ is placed at the cache of SBS. Otherwise, $xni$= 0. The MBS multicast transmission of a file $\imath$ will happen if a requester cannot detect $\imath$ in aanSBS. This happens when either the request for file $\imath$ is generated within an area that is not in the coverage area of any of the SBSs $n0 \in r$, or the request for file $\imath$ is generated within the coverage area of a SBan $n \in r$. Now we express with $J(x)$as the energy cost for servicing all the demands for $\imath$, which depends on the caching policy $x$. Concerning $\mathcal{r}$ that produces requests $\imath$ within a tiperiodan individual MBS multicast transmission of cost $Cb + Cw$ occurs. If a requester cannot find $\imath$ in the accessed SBS, $Yni$ = 1; the term $Yni$ is used to indicate whether the MBS multicast transmission will occur. In the other case, $Yni$ = 0, which indicates that all the requests are satisfied by the accessed SBSs, where the requests in area $\mathcal{n}$ incur cost $Cn$. Further, the following expression indicates that the number of the requests for file $\imath$ in a coverage area follows the Zipf distribution, and $\lambda ni$ is the rate parameter:

$$pnid = 1 - e{-}\lambda nid \qquad (1)$$

Thus, the problem that minimizes the total servicing cost by single multicast transmission is written as follow:

$$J_{ni} = \sum r q(Y_{ni}(C_b + C_w) + (1 - Y_{ni}) \sum n \in r C_n) \qquad (2)$$

However, the multicast caching problem is an NP-hard problem as illustrated in [1]. Because of that, we present the approach algorithm in the next section that finds a solution to the multicast caching problem greedily.

## IV. HEURISTIC GREEDY ALGORITHM

To determine the files that should be cached in SBSs, we implement a heuristic greedy multicast algorithm to gain the optimal caching policy to minimize the cost incurred for serving the file requests of mobile users. This algorithm terminates when all the caches become full to ensure that no more files will be placed at this cache and it will start with empty caches. In more detail, ($In$) is the number of files already placed in the cache of SBS $n$ of the algorithm. ($D$) is the set of all the pairs ($n$, $i$) for the file $i$ if it is not placed in the cache. When the caches $n*$ become full, the algorithm will exclude all the pairs ($n*$, $i$)

## V. CONTENT POPULARITY: ZIPF LAW

It has been empirically revealed that the popularity distribution of a set of items on the web can be approximated using Zipf relative popularity distribution. The parameter α of the Zipf distribution thus defines how heavy-tailed is the distribution. For example, in our case, it means how much the relative popularity between content items varies. Simulating a system of network caches, randomly generated data based on a Zipf distribution can be an alternative to a real network traffic trace that can be difficult to achieve. It has been proposed in several works that only a small portion of the whole content, termed popular content, is reached by a large fraction of users, while a large amount of content remains unpopular and sparsely requested. This allows one to focus on the library of popular files, where each file is endowed with a popularity distribution, which is empirically determined to follow Zipf law. Where N is the number of elements, k is their rank and s is the value of the exponent characterizing the distribution [32]. Then, the Zipf law predicts the frequency of elements of rank k,f (k, s, N) out of the population of N elements [31]:

$$f\,(k,s,N\,) = \frac{1/K^{S}}{\sum_{n=1}^{N}(\frac{1}{n^{S}})} \qquad (3)$$

we assumed that we have 1000 video files associated with (x) number of requests for each. Thus, we apply Zipf law on them so that we have the popularity of the (frequency) of each. If a new file (x) is requested, the popularity will be updated, such that the file (x) rank "popularity" is determined. Assuming that we have 6 SBSs, and 100 Users on each, the maximum requests (when all users request the same file) will be 600. We sorted the files in a table by its popular demand, and it can always update its rank based on the new user request. The Zipf model has been used to examine the cumulative distribution of requests to popular documents in each trace as illustrated in Equation (3) where (n) is the maximum number of requests and (s) is the value that characterized the contribution (0.8).

## VI. PERFORMANCE EVALUATION

To decide the cache placement we implement the proposed algorithm when the operator employs single multicast transmission for the demands of the same file in the same period (MC-MT). Next, we examine the above scheme with the performance of the unicast transmission (PC-UT) which is known as a conventional mode applied in the most caching scheme. Every SBS stores in its cache the popular files locally and separately from the others. Thus, every user demand is served by a separate unicast transmission. Please see Fig. 4.

To evaluate the heuristic algorithm, we examine the energy cost achieved by the above schemes as a function of the cache sizes. We found that PC-UT effects the highest servicing cost compared to the other MC-MT scheme. This is because the MC-MT scheme serves several aggregated requests through a single multicast rather than many unicast transmissions. Moreover, increasing the cache sizes decreases the servicing cost of the operator as more requests are served locally without the cooperation of the MBS. Please see Fig. 5

Finally, in Fig.6 we show the cumulative probability of access for top r% of documents in each trace. As shown clearly, the top 1% of the documents account for about 13% of all requests, and the top 2% of the documents account for about 7% of all requests.

## VII. CONCLUSION AND FUTURE WORK

This paper proposed to start by caching the files that have the highest probability to be requested by the users and depending on the preferences of users that are connected to each SBS. Over time, the popularity of files differs from one SBS to another SBS. Interestingly, we find that increasing the cache sizes decreases the servicing cost of the operator as more requests are served locally without the cooperation of the MBS. Further, the performance of the heuristic greedy multicast caching algorithm shows effects lowest servicing cost compared to unicast caching.

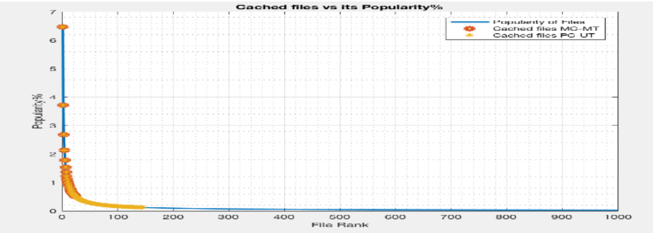

**Figure 4: The cached file**

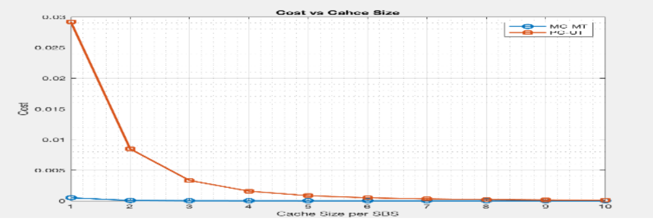

**Figure 5: The Performance comparison of PC-UT and MC-MT for the cache size**

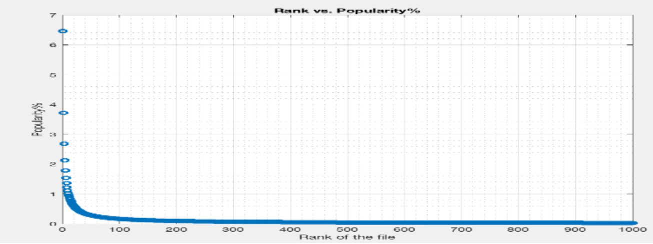

**Figure 6: The most popular files**

## REFERENCES

[1] Cisco,"Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021" White Paper "http://www.cisco.com/en/US/netsol/ns827/networking_ solutions_white_papers_list.html," March, 28. 2017.

[2] J. *Wannstrom, K. Mallinson,W. Harbor,* “HetNet/Small Cells”, 3GPP, “http://www .3gpp.org/hetnet”

[3] K. Hamidouche, W. Saad, and M. Debbah, “Many-to-many matching games for proactive social-caching in wireless small cell networks,” in 12th International Sym- posium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pp. 569–574, May 2014.

[4] small Cell Forum, ”Small cells - what’s the big idea?”, Feb. 2012.

[5] D. Chambers, Think Small Cell, “Small Cell Backhaul“ [online]“ https://www.thinksmallcell.com/Backhaul/data-caching-reduces-backhaul-costs-for-small cells-and-wi-fi.html” , May 2013.

[6] A. Sengupta, R. Tandon and T. C. Clanc, "Layered caching for heterogeneous storage," 2016 50th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2016, pp. 719-723.

[7] L. A. Belady, “A study of replacement algorithms for a virtual-storage computer,” IBM Syst. J., vol. 5, no. 2, p. 78–101, 1966.

[8] J. Wang, “A survey of web caching schemes for the internet,” ACM SIGCOMM Computer Communication Review, vol. 29, no. 5, pp. 36–46, October 1999.

[9] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, “A survey of information-centric networking,” IEEE Communications Magazine, vol. 50, no. 7, pp. 26–36, 2012.

[10] M. R. Korupolu, C. G. Plaxton, and R. Rajaraman, “Placement algorithms for hierarchical cooperative caching,” Journal of Algorithms, vol. 38, no. 1, pp. 260 – 302, 2001.

[11] K.C.AlmerothandM.H.Ammar,“The Use of Multicast Delivery to Provide a Scalable and Interactive Video-on-Demand Service,” Selected Areas in Communications, IEEE Journal on, vol. 14, no. 6, pp. 1110–1122, 1996.

[12] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” IEEE Transactions on Information Theory, vol. 60, no. 5, pp. 2856–2867, May 2014.

[13] M. A. Maddah-Ali and U. Niesen, “Decentralized coded caching attains order-optimal memory-rate tradeoff,” IEEE/ACM Transactions on Networking, vol. 23, no. 4, pp. 1029–1040, Aug. 2015

[14] S. Woo, E. Jeong, S. Park, “Comparison of caching strategies in modern cellular backhaul networks,” June, 2013.

[15] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, G. Caire, ”FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers,” Proceedings of the IEEE Infocom, pp. 1107-1115, 2012.

[16] E. Bastug, M. Bennis and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," in IEEE Communications Magazine, vol. 52, no. 8, pp. 82-89, 2014.

[17] D. Liu, C.Yang, “Cache-enabled Heterogeneous Cellular Networks: Comparison and Tradeoffs” Feb, 2016.

[18] S. S. Bidokhti, M. Wigger, and R. Timo, “Erasure broadcast networks with receiver caching,”in*Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 1819–1823, July 2016.

[19] R. Tandon and O. Simeone, “Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks,” in *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 2029–2033, July 2016.

[20] C. Yang, Z. Chen, Y. Yao, and B. Xia, “Performance analysis of wireless heterogeneous networks with pushing and caching,” in *Proc. IEEE ICC*, 2015.

[21] S. H. Chae, J. Y. Ryu, T. Q. S. Quek, and W. Choi, “Cooperative transmission via caching helpers,” in *Proc. IEEE GLOBECOM*, 2015.

[22] ]M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, “Content-aware user clustering and caching in wireless small cell networks,” in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, Aug 2014, pp. 945–949.

[23] S. Mu ̈ller, O. Atan, M. van der Schaar, and A. Klein, “Smart caching in wireless small cell networks via contextual multi-armed bandits,” in *IEEE Int’l Conference on Communications (ICC)*, June 2016.

[24] E. Ba ̧stug ̆, M. Bennis, and M. Debbah, “Anticipatory caching in small cell networks:A transfer learning approach,” in 1st KuVS Workshop on Anticipatory Networks, Stuttgart, Germany, 09/2014 2014.

[25] E. Ba ̧stug ̆, K. Hamidouche, W. Saad, and M. Debbah, “Centrality-based caching for mobile wireless networks,” in 1st KuVS Workshop on Anticipatory Networks, Stuttgart, Germany, 09/2014 2014.

[26] S. Chan and F. A. Tobagi, “Scalable services for video-on-demand,” Stanford, CA, USA, Tech. Rep., 1998.

[27] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, “Web caching and zipf like distributions: evidence and implications,” in *INFOCOM ’99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1, Mar 1999, pp. 126–134 vol.1.

[28] M. M. Amiri, Q. Yang, and D. Gunduz, “Coded caching for a large number of users,” CoRR, vol. abs/1605.01993, 2016.

[29] J. Hachem, N. Karamchandani, and S. Diggavi, “Content caching and delivery over heterogeneous wireless networks,” in 2015 IEEE Conference on Computer Communications (INFOCOM), April 2015, pp. 756–764.

[30] K.Poularakis,G.Iosifidis,V.Sourlas,andL.Tassiulas,“Exploiting caching and multicast for 5g wireless networks,” IEEE Transactions on Wireless Communications, vol. 15, no. 4, pp. 2995–3007, April 2016.

[31] Zipf's law. Wikipedia, Available from: https://en.wikipedia.org/wiki/Zipf%27s_l

[32] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, “Web caching and zipf like distributions: evidence and implications,” in *INFOCOM ’99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1, Mar 1999, pp. 126–134 vol.1