AN ABSTRACT OF THE THESIS OF

Ragda Abuhadra for the degree of Master of Science in Electrical and Computer Engineering presented on August 30, 2017.

Title: Hybrid, Proactive In-Network Caching for Mobile On-Demand Video Streaming

Abstract approved: ____

Bechir Hamdaoui

Mobile video streaming has become an essential application in mobile wireless networks, making up most of the mobile data of today's Internet traffic. Studies have shown that mobile video data is projected to make up about 78 percent of the global mobile data traffic, and that global mobile data traffic is expected to increase sevenfold by 2021. Massive small cell base station (SBS) deployments have emerged as a potential solution promising to fulfill these unprecedented mobile data demands, by offering great coverage enhancements and maintaining high quality of video streaming. However, due to relatively small cell sizes and high user mobility, mobile video streaming in dense SBS networks faces fundamental challenges such as intermittent connectivity and frequent handoffs, causing degradation in video streaming quality. In this thesis, we tackle this issue by introducing a hybrid proactive in-network caching framework that stores some popular videos at the edge of the network, namely at the SBSs, while also pre-caching video contents in advance to better service mobile users. The proposed framework essentially reduces the need for bringing every requested video from the core (original) network, which results in alleviating network congestion by reducing back-haul traffic

and in improving mobile video streaming experience by avoiding service discontinuity during handoffs. We develop a simulation framework using MATLAB to study the performance of the proposed hybrid proactive caching technique, and show using simulations that the proposed technique can effectively improve video quality of experience and reduce back-haul traffic.

Hybrid, Proactive In-Network Caching for Mobile On-Demand Video Streaming

by

Ragda Abuhadra

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented August 30, 2017
Commencement June 2018

Master of Science thesis of Ragda Abuhadra presented on August 30, 2017.

APPROVED:

Major Professor, representing Electrical and Computer Engineering

---

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Ragda Abuhadra, Author

ACKNOWLEDGEMENTS

After thanking Allah, I would like to thank my husband and my parents for believing in me and for being with me throughout the way. I would also like to thank my academic advisor, Prof. Bechir Hamdaoui for his counseling and guidance. Last but not least I would like to thank all my friends and colleagues who supported me in my academic journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

Global mobile data traffic has been increasing substantially over the last several years, and is projected to reach a data volume of 49 Exabytes per month by 2021, of which 78 percent will be mobile video data [1]. Therefore, one of the greatest challenges future wireless networks (e.g. 5G mobile networks) are faced with is to accommodate and keep up with these overwhelming data demands. One of the potential solutions that emerged as a way of fulfilling these unprecedented traffic demands and increasing network capacity is through deployments of small cell networks (SCNs) [2, 3], also often referred to in the literature as femtocells/picocells [4, 5]. Small cell networking represents a novel networking paradigm based on the idea of deploying short-range, low-power, and low-cost small base stations (SBSs) (e.g., picocells, femtocells with a range of 200 meters and under) to provide high data rates and increased network capacity [6–8]. The problem with this paradigm, however, is that users with medium to high mobility can suffer severely from frequent handoffs due to these small cell deployments, leading to (possibly severe) degradation in the quality of service of ongoing communications [9–14]. As such, the support of mobile video streaming in dense small cell networks presents several challenges. Since mobile users can have very short connection times to the base station and thus can experience frequent handovers across subsequent base stations when they are on the move, there is a need to frequently rebuild the route to remote video servers through different base stations, which can incur substantial communication overheads and long content delivery latencies. The latency for re-building the route to deliver video content from the remote video server has a significant impact on the video streaming quality, since high resolution videos require high data rate connections with little to no disruption [14,15]. Such an obstacle drove the researchers to investigate the possibility of bringing the contents closer to the end-users to make it possible to reduce network traffic and improve quality of experience simultaneously. Therefore, the concept of living at the edge emerges as a result of such efforts, which essentially means proactively caching most popular contents at the base stations [16–20] so that content is stored closer to end users. This concept is also referred to in the literature as fog or edge cloud computing [21–25].

In this thesis, in order to provide a satisfactory quality of mobile video streaming through dense small-cell network deployments, a hybrid, proactive in-network caching framework is proposed. In this framework, each base station is equipped with a cache memory that is divided into two parts. One part is designated for proactively caching the most popular content using one of the caching protocols, and the other part is designated for pre-fetching ahead video content from the remote video server to serve upcoming mobile users. This system assures that a video content is immediately delivered to the mobile user once it enters a new cell, which mitigates the delay produced from frequent route re-building to a remote video server and hence reduces content access delay. It also reduces the use of the back-haul network, thereby saving back-haul bandwidth. To this end, our contributions in this thesis are:

- We evaluate the performance of two existing proactive caching schemes.
- We propose an efficient hybrid proactive caching framework that improves content delivery latency and alleviates network traffic congestion.
- We evaluate our proposed framework and illustrate how it achieves a satisfied trade-off between back-haul network load and user's quality of experience.

The remainder of this thesis is structured as follows. Literature review is presented in Chapter 2. In Chapter 3, we present our system model. In Chapter 4, we describe our proposed scheme. In Chapter 5, we present our simulation framework and performance evaluation. Finally, we present our conclusion and future works in Chapter 6 and 7 respectively.

Chapter 2: Literature Review

2.1 Small Cell Networks

As data traffic load is increasing and subscribers' performance expectations are rising, it is well understood that supplementing macro networks with small cells is an effective way to increase coverage and improve spatial reuse, boost capacity, and offload traffic more efficiently and hence improve users' quality of experience (QoE) [6]. Small cell network (SCN) means smaller cell size (< 200 meters), higher cell densification (a fraction of a square meter), making network access closer to end users [8].

However, deploying additional network nodes indoors and outdoors within a local-area range to create small cell networks can give rise to a number of challenges, such as dealing with higher control traffic, interference, and backhaul congestion. Therefore, network optimization processes will be essential. Among other literature, [3, 26] present a comprehensive survey about SCNs and femtocells respectively. [3] tackles various aspects from interference management, cell association, stochastic network modeling and mobility management. In [27], E. Bakin, et al analyse picocell network capacity with dominating video streaming traffic by proposing one model and two techniques of congestion probability estimation in centralized wireless networks. In [28, 29], Chakchouk et al. provide a cross-layer analysis of uplink performance in femtocell networks by characterizing the uplink physical interference and studying its impact on the delay and traffic loss rate.

2.2 Video Streaming

Video or multimedia streaming is the mechanism in which media is sent as a compressed content in a continuous data stream and is played as it arrives [30,31]. In other words, the user does not need to wait to download the whole file to play it. In addition, on-demand video streaming allows users to select and watch video content at their own chosen time, rather than having to watch at a particular broadcast time. As mentioned before, video data constitutes a significant fraction of today's Internet traffic and has rapid growth in mobile networks. This increasing demand has introduced several challenges in delivering videos with a high quality of experience to the users over wireless networks (e.g. wireless sensor networks [32–34], mobile wireless networks [35], multi-hop wireless networks [36–38], vehicular networks [39], and MIMO-enabled wireless networks [40, 41]). These challenges range from limited wireless spectrum/bandwidth resources to user mobility and handoffs, and from interference management to resource allocation and optimization. As a result, quality of user's experience for mobile video streaming is of critical concern today. Among other literature, [42] discusses such challenges and presents various solutions such as rate adaptation and scheduling, and video multicasting when a group of users in a particular area are interested in the same video at the same time. Authors of [43] discuss QoE measurements for Video Streaming over wireless networks. In [44] authors present a smart base station-assisted partial-flow device-to-device offloading system in order to alleviate the cellular network traffic load by effectively offloading parts of the video traffic to D2D networks which provides seamless video streaming services to users. In [45], a new protocol called Enhanced User Datagram Protocol (EUDP) for video streaming in VANET is proposed. Unlike User Datagram Protocol (UDP), this protocol uses Sub-Packet Forward Error Correction (SPFEC), and adopts the unequal protection of video frame types in order to improve the video streaming quality. [46, 47] investigate QoE of video streaming over VANET.

Simultaneously others are investigating the possibility of bringing certain contents closer to the end users to reduce both the expected time for content delivery and the traffic over the links connecting the wireless network to the Internet backbone; this is called *proactive caching*.

2.3 Proactive Caching

Proactive caching emerges as one of the potential approaches used to enhance the quality of user experience (QoE) of mobile video streaming in wireless networks. As such, its impact on delay and capacity has been intensively investigated [16–20, 48–52]. In [16], authors explore proactive caching by proposing a mechanism of proactively caching files during off-peak periods based on file popularity and correlations among user and file patterns in order to alleviate back-haul congestion. Also, they propose a procedure that exploits the social structure of the network and leverages D2D communication. In [53], a proactive caching procedure using perfect knowledge of content popularity is studied. Two new caching policies based on the user preference profile are proposed in [54] and a video scheduling approach is developed which allocates the radio access network back-haul resources to the video requests so as to reduce video latency and increase network capacity. A FemtoCaching architecture relying on cache enabled user devices and small base stations is introduced in [55] and a wireless distributed helper system is proposed in [56] as a promising way of reducing the bottlenecks in wireless video delivery. Most of the existing works consider classic proactive caching by fetching and storing only popular contents in the cache memory of the base station proactively. In these, depending on the cache size and caching protocol, a particular number of popular files are stored in the base station memory to achieve high hit ratios and hence reduce the back-haul transmissions whenever the requested files exist in the cache, thus reducing delay [48,54]. D. Liu and C. Yang in [57] show that energy efficiency can be improved by introducing cache at the pico base stations and if the file catalog size is not too large. However, when the desired files are not present in the cache, content delivery delay will increase due to the need for connecting to the remote server to download the requested file. Authors of [58] investigate the proactive caching strategy for maximizing the average QoE of scalable video coding over small cell networks. They prove that caching videos in collocated small cell base stations may not only reduce transmission range, but also bring channel diversity gains. Security is also a challenge that is facing video streaming in cache-enabled small cell networks, where some of the cache-enabled SBSs helping in video delivery might be untrusted. [59] addresses this issue and proposes joint caching and scalable video coding of video files for improving system performance.

In [19, 20], another proactive caching scheme is adopted to distribute video content from remote video servers to the candidate base stations ahead of the user association in high-mobility environments, so that the video content required by a mobile user is immediately available when the user moves to the new cell. In this case for every content request of every user, the base station has to use the back-haul connection to deliver the required content since there are not any pre-stored popular contents in the cache memory. This approach is likely to increase congestion, back-haul transmission, and initial connection delay. Therefore, there is a necessity to consider a tradeoff between the cache hit ratio and the video frozenness duration. In [17, 18], the authors propose a content-centric caching approach for urban, highly dense cities, and use LinkNYC [60] as a use-case to evaluate and validate their proposed caching framework. Their proposed caching framework relies on content popularity, population densities, and cooperation among in-network devices (referred to as cloudlets) to bring content as close to end-user as possible to improve overall network performance by avoiding to traverse the Internet for content as much as possible, thereby reducing infrastructure and bandwidth costs while improving content delivery responsiveness and users' QoE in general.

In this thesis, we are considering adopting the scheme in [19] and adding a pre-stored popular content in the cache memory depending on the capacity constraint of the local memory. When a base station receives a content request from an upcoming user it first searches in its local stored content, and in the event the request does not match any of the stored contents, the SBS requests it from the remote core network.

Chapter 3: System Model

In this thesis, a three-tier network hierarchy is considered as shown in Figure 3.1. The hierarchy is composed of a core network containing a library of files, small base stations (SBSs) equipped with local caches, and a set of mobile users.

We consider a scenario of a set of N users, $\mathcal{U}= [U_1, U_2, .., U_n, ..U_N]$, moving in one direction on a segment of a highway. The highway is covered by a set of M low power SBSs, $\mathcal{B}= [B_1, B_2, .., B_m, ..B_M]$, covering a set of M small cells, $\mathcal{C}= [C_1, C_2, .., C_m, ..C_M]$. Each base station B_m covering a cell C_m has a local cache memory with size L (L is expressed in terms of the percentage of the total library content size). This local cache memory is divided into two parts. Part-I of size μL ($0 <\mu< 1$), containing a catalogue of a set of popular files $\mathcal{F}= [F_1, F_2, .., F_k, ..F_K]$, ranking from 1 (the most popular) to K (the least popular) based on popularity and following Zipf distribution with parameter s=1 [61]. Part-II is of size (1-μ)L, and designated for storing parts of the requested contents of the upcoming users. We assume that all local caches are of the same size and have the same popular contents which are updated periodically.

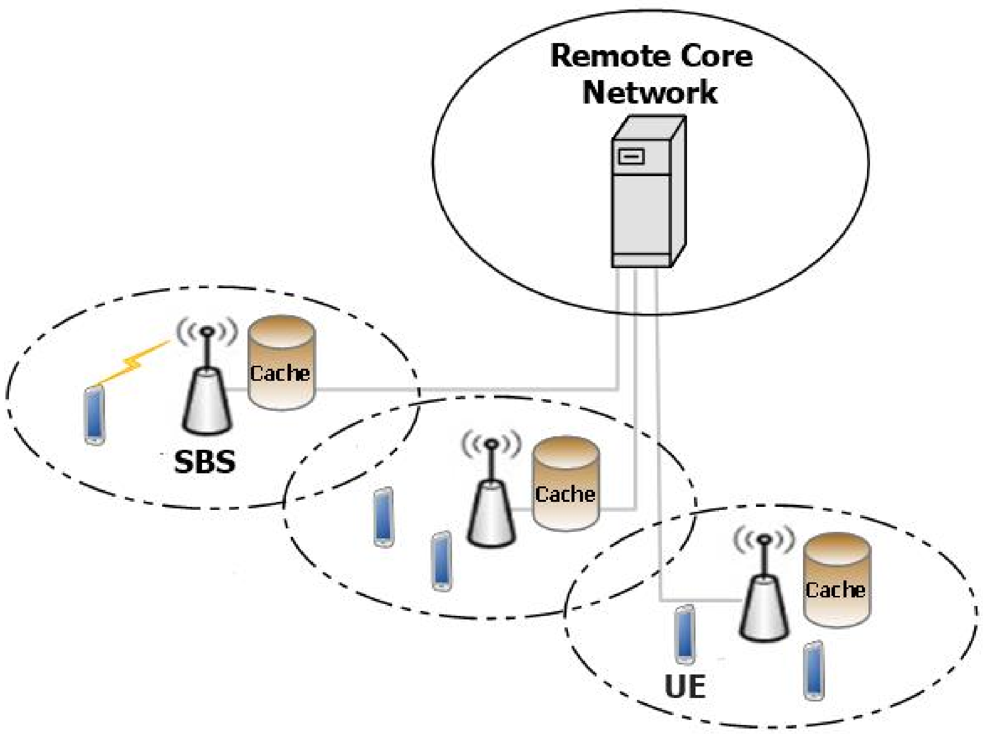

Figure 3.1: System Model

Chapter 4: Hybrid, Proactive In-Network Caching

Video content is segmented into blocks and different users' video blocks pre-loaded in the local cache of each base station are stored based on the order in which the users become associated with the base station. We consider that the number of SBSs M to be large enough to offer full coverage of the highway; that is, mobile users do not exceed the area on the highway covered by the base station set $\mathcal{B}$ within the considered time duration. Since users move in one direction along the highway, when user U_n is currently associated with base station B_m, the next associated base station is predicted to be B_{m+1}. When U_n initially connects to base station B_m and sends a content request, B_m first checks whether Context(U_n) is available in its local cache, and if it is, it pre-fetches the size of $S_{n,m}$ and delivers it to the user. Otherwise; i.e., if it is not available, it requests Context(U_n) from the remote video server. Before a user U_n moves into a new cell C_{m+1}, depending on the amount of available space in B_{m+1}'s local cache and the time the user will spend in the C_{m+1} cell, the system can determine the specific size of video content, $S_{n,m+1}$, to be cached ahead, and delivered once user U_n connects to base station B_{m+1}.

Since each B_m is equipped with a local cache memory whose Part-II size is (1-μ)L, we have

$$\sum_{n=1}^{N} \widetilde{S}_{n,m+1} \leq (1-\mu)L \tag{4.1}$$

where $\widetilde{S}_{n,m+1}$ is the local Part-II memory size allocated by base station B_{m+1} to user U_n. $\widetilde{S}_{n,m+1} = 0$ when Part-II space in the local cache of B_{m+1} is fully occupied by other users and there is no available memory space to be allocated to user U_n.

We assume that each user requests a video file of the size S, where the arrival of the event that the user U_n requests a file F_i follows a Poison distribution with the rate of t_n. Each user U_n stays in the cell C_m for time duration $T_{n,m}$ until it moves into the next cell C_{k+1}. This time duration is affected by the length of the highway located in the cell C_k and the speed V_n of the mobile user and it is independent of the previous visited cells and the previous length of video streaming duration. $T_{n,m}$ is assumed to be larger than the time needed to deliver the required content to B_{m+1}.

Chapter 5: Simulation Framework and Performance Evaluation

5.1 Simulation Framework

In this chapter, we conduct extensive MATLAB simulation to investigate the performance of the proposed caching scheme and compare it to the classic scheme as well as the on-the-fly scheme (both schemes are described next). Multiple cache sizes L and different values of cache fraction μ are tested. The numerical results are conducted for 50 mobile users moving in a highway segment covered by a series of 70 SBSs, each with a 100-meter range. We use circle cells for mathematical tractability. The system parameter settings are summarized in Table 5.1.

Parameter	**Value**
Number of Mobile Users (N)	50
Number of SBSs (M)	70
Cell Size	100 m
Mobile User Speed (V_n)	75 mph
Number of Video Files	100
Size of Video File (S)	30 Gbit
Bit rate of Video File	0.5 Gbit/s
Local Cache Size (L)	20%, 30%, 40%
Cache Fraction (μ)	0.1, 0.3, 0.5, 0.7, 0.9
Utility Weight Parameter (α)	0.2, 0.5, 0.8
Back-haul Capacity	10 Gbit/s
Wireless Small Cell Link Capacity	1 Gbit/s
Delay for Remote Connection	25-50 ms

Table 5.1: Simulation Parameters

Remark. *In order to implement the proposed hybrid proactive caching method, a central controller is necessary. This controller is connected to all the SBSs to collect the updated data (i.e., the required video content, the movement status of each user, and the status of the local caches).*

5.1.1 Performance Metrics

This thesis investigates the impact of hybrid proactive caching in dense small-cell networks on back-haul network transmission overhead, which is defined as the percentage of back-haul bandwidth savings and captured by the cache hit ratio:

$$\gamma = \frac{\text{number of cache hits}}{\text{total number of video requests}} \tag{5.1}$$

where a hit occurs when the requested file F_i by user U_n exists in the local cache of B_m that covers C_m which user U_n is associated with.

The thesis also investigates the impact of the proposed scheme on video streaming quality (i.e., delay and video stalling), which is defined as the percentage of freezing time (η) of the total streaming duration.

In our work, we are finding the cache size L and the value of the cache fraction μ that maximizes the following utility function, (Υ):

$$\Upsilon = \alpha\bar{\eta} + (1-\alpha)\gamma \tag{5.2}$$

where $\bar{\eta} = 1 - \eta$ and α is a weighting value. This utility function captures both performance metrics, freezing time and hit ratios, and offers a way to balance, through the design parameter α, between these two conflicting objectives.

5.1.2 Baseline Approaches and Their Performance Tradeoffs

In this section, we compare the performance of the proposed scheme to two existing caching approaches: classic proactive caching [48] and on-the-fly proactive caching [19]. We next provide a brief description of each of these two approaches.

The classic proactive caching approach consists of providing the nearby base stations with local caches of a certain size L for storing some number of popular files in order to reduce the usage of the back-haul network and to reduce latency when the requested files exist in the local caches. The on-the-fly caching approach, on the other hand, follows the same concept except the local caches do not contain any popular files. The cache space of size L is only used to store some contents' parts ahead to be pre-fetched later. The main goal here is to make the contents of the requested files ready at the future base stations before the users' arrivals so as to reduce latency.

In order to motivate the need for the proposed hybrid caching approach, we now present some results illustrating the performance tradeoffs between these two existing approaches. As it can be noticed from Figure 5.1, the on-the-fly proactive caching scheme improves the user experience. Video streaming under this on-the-fly caching suffers less installing duration when compared to the classic one. The first scheme eliminates delay effect from streaming duration caused by establishing the connection to the remote core network each time a user enters a new cell and requests a video by pre-loading the video content to the upcoming base stations ahead of the user's arrival. Yet, in the classic scheme, if the requested video is not available in the cache, the SBS needs to establish a connection to the remote core network to download the file.

On the other hand, the classic proactive caching performs better than the on-the-fly scheme when it comes to the back-haul bandwidth savings; this is illustrated in Figure

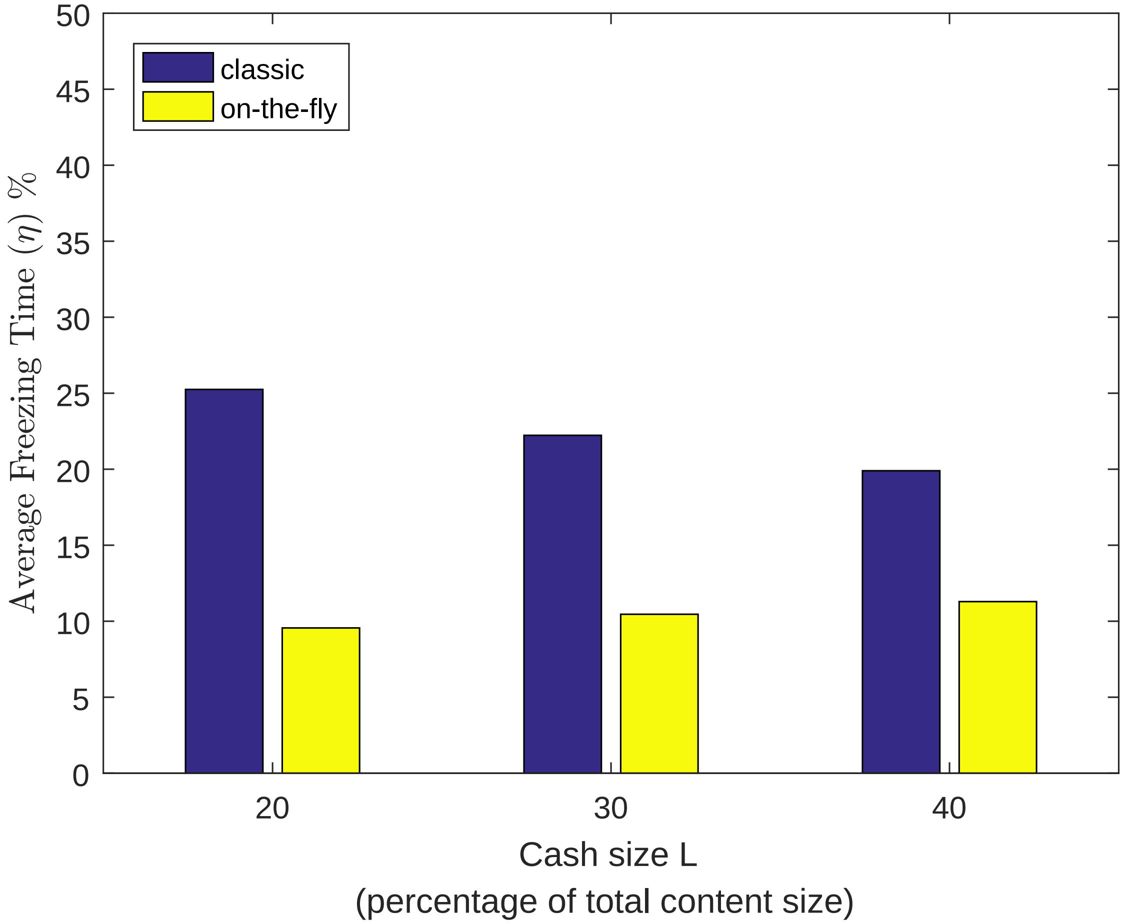

Figure 5.1: Average Freezing Time Comparison

5.2. In on-the-fly scheme, the content of every single video request should be downloaded from the remote core network, which increases the usage of the back-haul network and causes high traffic congestions, whereas in the classic scheme, the back-haul network is used only if the requested content does not exist in the local cache. It is also noticed that the chances for getting the video from the local cache increases with the cache size, since more files are stored. Figures 5.1 and 5.2 show the need for making a tradeoff between the user's QoE and the amount of traffic load put on the back-haul connection. This can be achieved by merging the features of the previous two schemes. Our proposed hybrid technique is centered around such an idea.

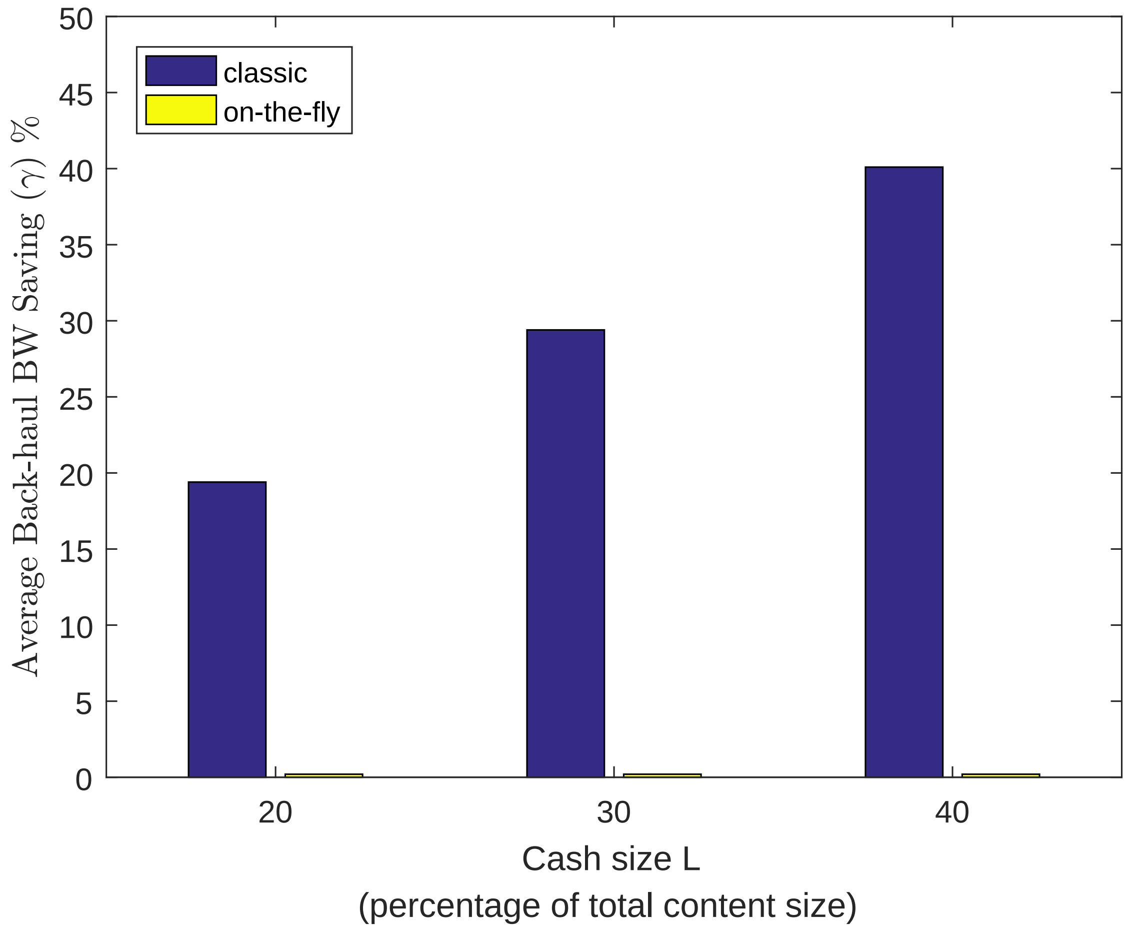

Figure 5.2: Average Back-haul BW Saving Comparison

5.2 Result Analysis

We now present the performance of our proposed hybrid proactive caching scheme and compare it to those achieved under the other two caching schemes. We also show how it can successfully achieve a good balance between having a better user's QoE and lesser back-haul network load. In our analysis, we consider testing the performance metrics:

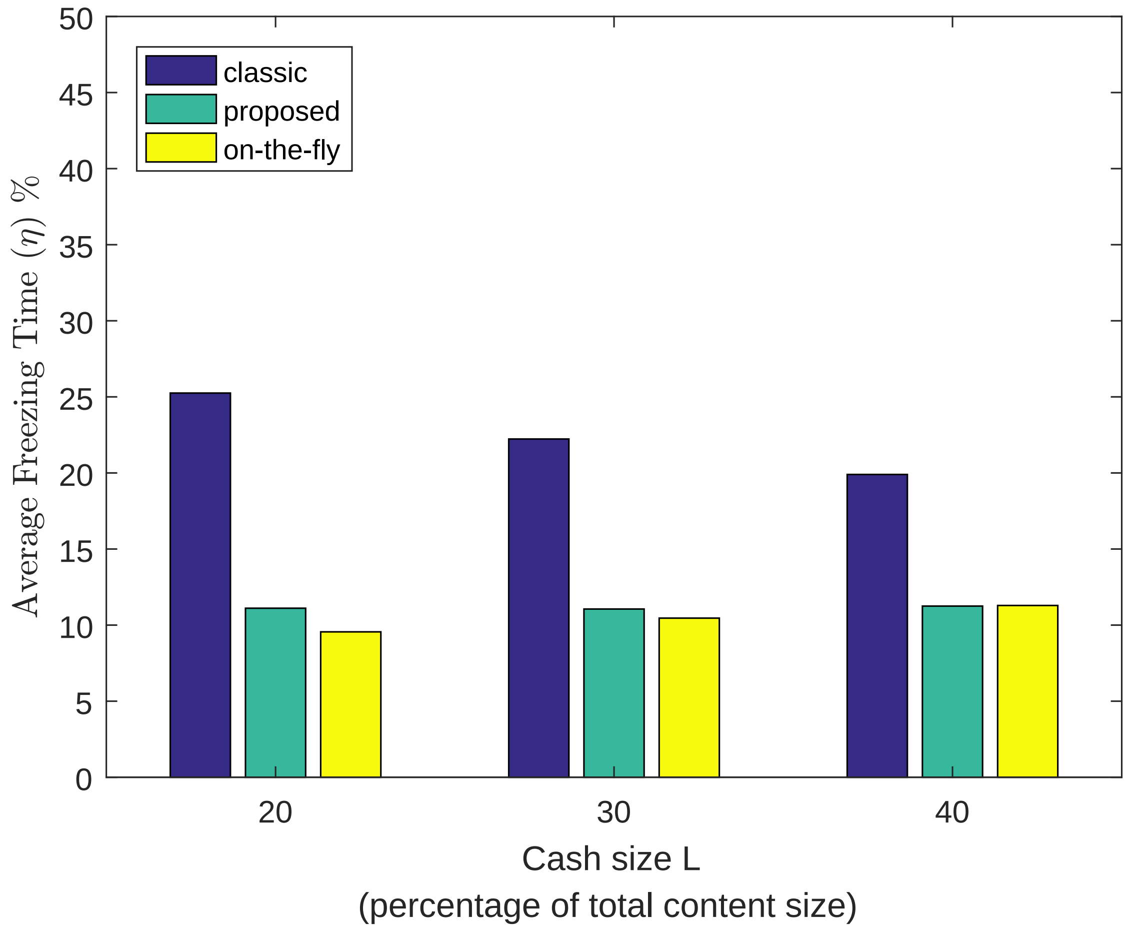

Figure 5.3: Average Freezing Time Comparison

($\mu = 0.7$)

back-haul bandwidth saving (γ) and freezing time percentage (η). It is observed that the hybrid proactive caching scheme outperforms the classic caching scheme vis-a-vis of user's QoE, which is reflected in freezing time reduction. Concurrently the proposed

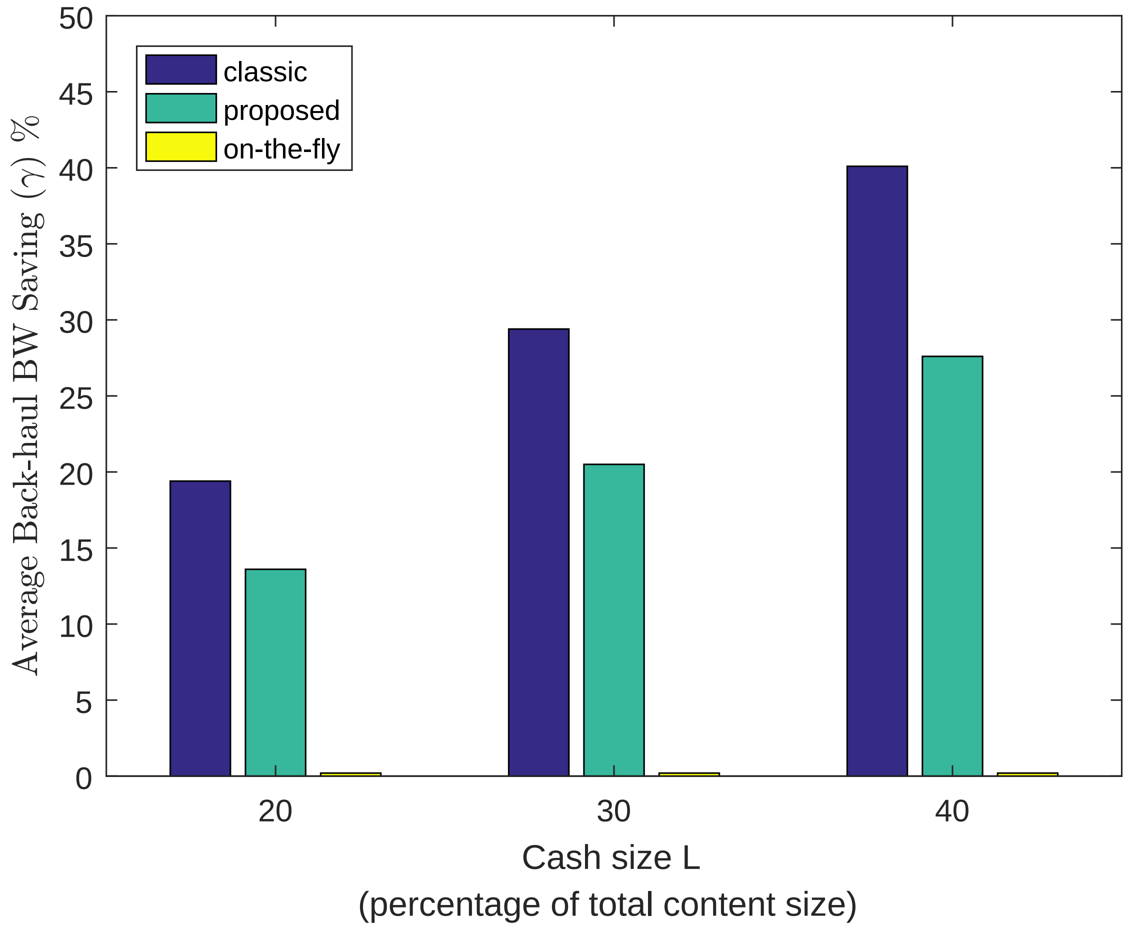

Figure 5.4: Average Back-haul BW Saving Comparison

($\mu = 0.7$)

scheme performs better than the on-the-fly caching in terms of back-haul bandwidth savings and congestion reduction at the back-haul link. These two resulted improvements can be observed in Figures 5.3 and 5.4, respectively. Figure 5.3 demonstrates that the proposed scheme reduces the average percentage of freezing time by 15% compared to the classic scheme. On the other hand, in Figure 5.4 the proposed scheme saves the back-haul bandwidth by nearly 30% while the on-the-fly scheme causes high load on the back haul network with 0% saving.

In our comparison, we also consider the utility function (Υ) (5.2), which combines the two performance metrics. For this, we test multiple weight values depending on the importance of each individual metric. Each of Figures 5.5 and 5.6 illustrates that for $\mu = 0.7$ (i.e. 70% of the local cache size stores popular contents while the other 30% is used to ahead pre-fetching), the proposed caching scheme achieves a higher value of Υ compared to the other two schemes, when α values are 0.5 and 0.8. It also can be observed in Figure 5.7 that when the local cache size L=40%, the proposed hybrid caching scheme achieves the highest Υ when $\alpha = 0.8$ since the cache size is large enough to accommodate the users' demand. This is due to the fact that Part-I of the cache contains enough number of popular files to achieve higher cache hit ratios. In addition, Part-II is large enough to serve most of the users simultaneously.

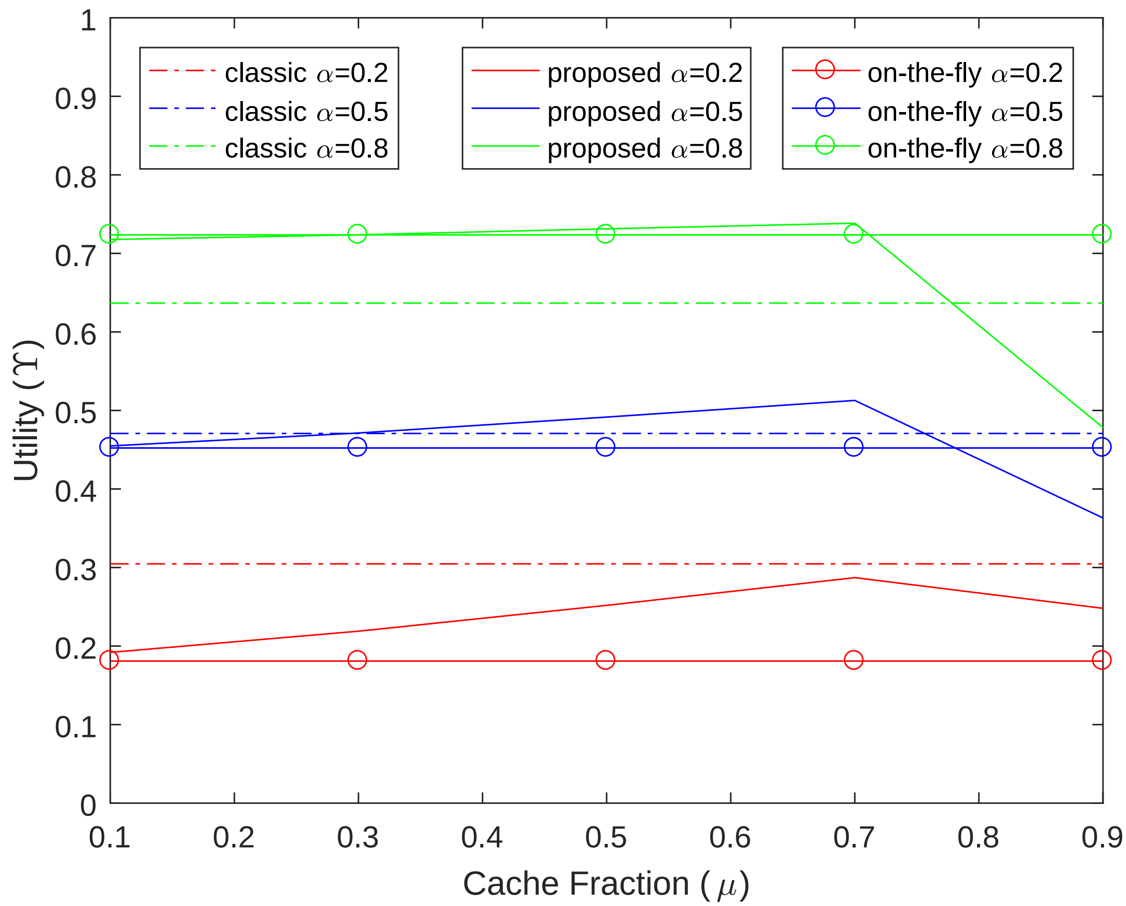

Figure 5.5: Trade off between user QoE and back-haul network load. (cache size L=20%)

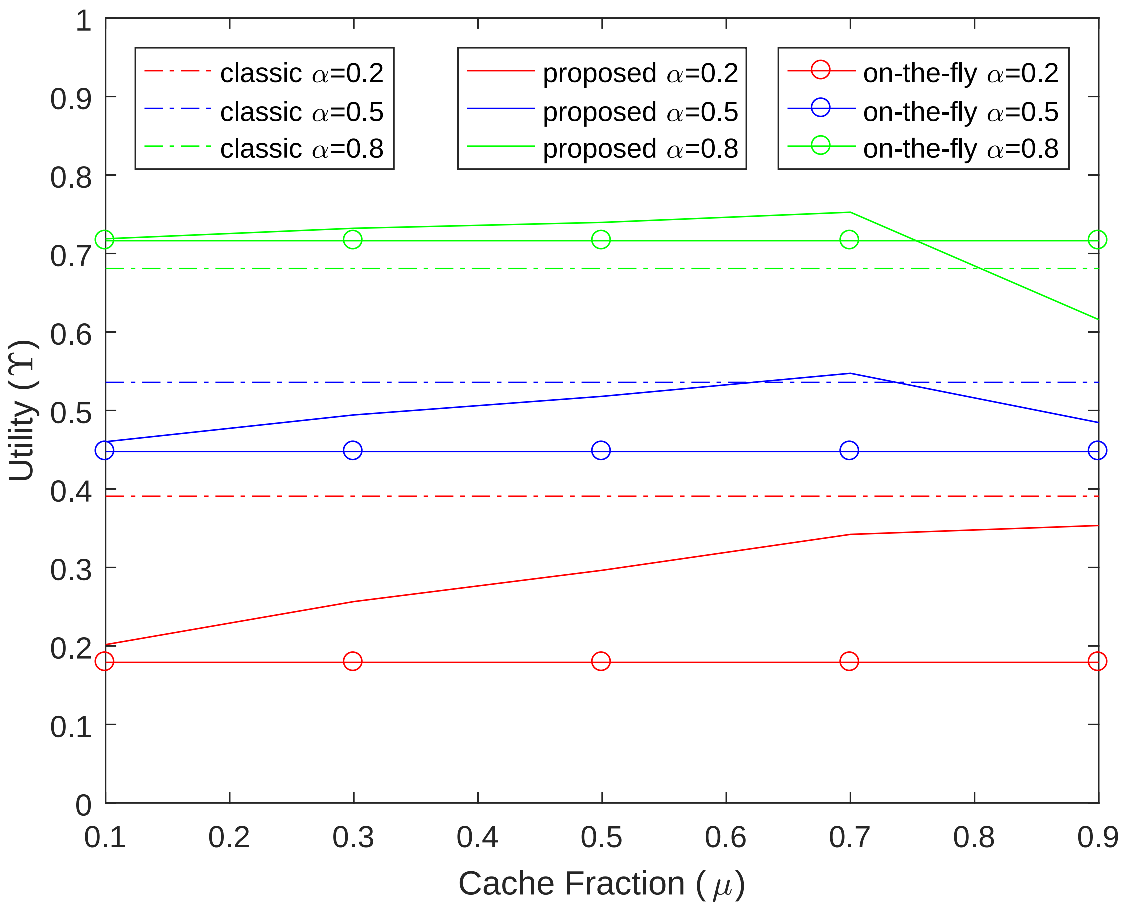

Figure 5.6: Trade off between user QoE and back-haul network load. (cache size L=30%)

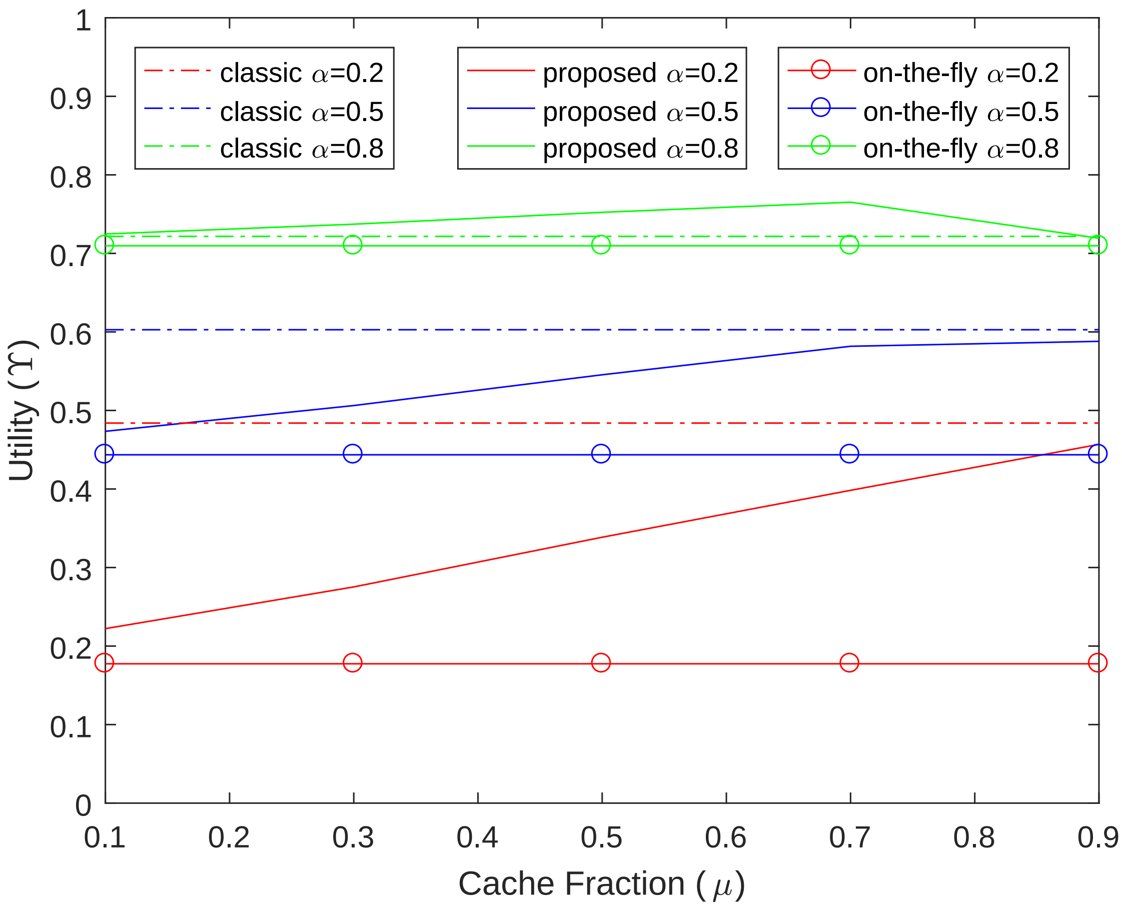

Figure 5.7: Trade off between user QoE and back-haul network load. (cache size L=40%)

Chapter 6: Conclusion

In this thesis, we have investigated the possibility of improving on-demand video streaming in dense small cell networks with high user mobility by proposing a hybrid proactive caching framework. The framework merges the concept of bringing the content closer to the user by caching some of the most favourite contents at the edge of the network with the concept of providing parts of the content ahead of the moving user at the upcoming base stations. We show that content delivery delay and back-haul traffic load can both be significantly reduced when using the proposed hybrid in-network caching approach.

Chapter 7: Future Work

I believe that there are several interesting problems that are worth investigating and exploring to further strengthen the results of this thesis. One of these problems is energy efficiency where the proposed scheme can be enhanced to account for energy consumption. Another problem is content distribution among local caches, i.e., considering storing different files' sets in adjacent SBSs caches with cooperative caching scheme deployment and investigate the effect of such a framework on cache hit ratio and content delivery time. Also, it could be interesting to consider other mobility scenarios and models that suit dense cities.

Bibliography

[1] Cisco visual netwroking index: Global mobile data traffic forecast update, 2016-2021. White Paper.

[2] J. G. Andrews, "Seven ways that hetnets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, March 2013.

[3] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2522–2545, Fourthquarter 2016.

[4] N. Chakchouk and B. Hamdaoui, "Qos-aware autonomous distributed power control in co-channel femtocell networks," in *Global Communications Conference (GLOBECOM), 2012 IEEE*. IEEE, 2012, pp. 567–571.

[5] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, 2012.

[6] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5g ultra-dense cellular networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72–79, February 2016.

[7] N. Chakchouk and B. Hamdaoui, "Estimation-based non-cooperative power allocation in two-tier femtocell networks," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*. IEEE, 2011, pp. 1–6.

[8] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, April 2012.

[9] H. Sinky and B. Hamdaoui, "Cross-layer assisted tcp for seamless handoff in heterogeneous mobile wireless systems," in *Globecom Workshops (GC Wkshps), 2013 IEEE*. IEEE, 2013, pp. 4982–4987.

[10] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications magazine*, vol. 46, no. 9, 2008.

[11] H. Sinky, B. Hamdaoui, and M. Guizani, "Handoff-aware cross-layer assisted multipath tcp for proactive congestion control in mobile heterogeneous wireless networks," in *Global Communications Conference (GLOBECOM), 2015 IEEE*. IEEE, 2015, pp. 1–7.

[12] J.-M. Moon and D.-H. Cho, "Novel handoff decision algorithm in hierarchical macro/femto-cell networks," in *Wireless Communications and Networking Conference (WCNC), 2010 IEEE*. IEEE, 2010, pp. 1–6.

[13] B. Hamdaoui and P. Ramanathan, "A network-layer soft handoff approach for mobile wireless ip-based systems," *IEEE Journal on selected Areas in Communications*, vol. 22, no. 4, pp. 630–642, 2004.

[14] H. Sinky, B. Hamdaoui, and M. Guizani, "Proactive multipath tcp for seamless handoff in heterogeneous wireless access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4754–4764, 2016.

[15] M. Xing, S. Xiang, and L. Cai, "A real-time adaptive algorithm for video streaming over multiple wireless access networks," *IEEE Journal on Selected Areas in communications*, vol. 32, no. 4, pp. 795–805, 2014.

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

[17] H. Sinky and B. Hamdaoui, "Cloudlet-aware mobile content delivery in wireless urban communication networks," in *Global Communications Conference (GLOBECOM), 2016 IEEE*. IEEE, 2016, pp. 1–7.

[18] H. Sinky, B. Khalfi, B. Hamdaoui, and A. Rayes, "Responsive content-centric delivery in large urban communication networks: A LinkNYC use-case," *IEEE Transactions on Wireless Communications (under revision)*, 20017.

[19] J. Qiao, Y. He, and X. S. Shen, "Proactive caching for mobile video streaming in millimeter wave 5g networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 7187–7198, Oct 2016.

[20] H. Farahat and H. S. Hassanein, "Proactive caching for producer mobility management in named data networks," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, June 2017, pp. 171–176.

[21] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Replisom: Disciplined tiny memory replication for massive iot devices in lte edge cloud," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 327–338, 2016.

[22] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, 2016.

[23] S. Abdelwahab and B. Hamdaoui, "Flocking virtual machines in quest for responsive iot cloud services," *arXiv preprint arXiv:1610.00289*, 2016.

[24] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[25] S. Abdelwahab and B. Hamdaoui, "Fogmq: A message broker system for enabling distributed, internet-scale iot applications over heterogeneous cloud platforms," *arXiv preprint arXiv:1610.00620*, 2016.

[26] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, September 2008.

[27] E. Bakin, A. Borisovskaya, and I. Pastushok, "Analysis of capacity of picocell with dominating video streaming traffic," in *Proceedings of 15th Conference of Open Innovations Association FRUCT*, April 2014, pp. 3–8.

[28] N. Chakchouk and B. Hamdaoui, "Statistical characterization of uplink interference in two-tier co-channel femtocell networks," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International*. IEEE, 2012, pp. 968–973.

[29] ——, "Uplink performance characterization and analysis of two-tier femtocell networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 9, pp. 4057–4068, Nov 2012.

[30] S. Misra, M. Reisslein, and G. Xue, "A survey of multimedia streaming in wireless sensor networks," *IEEE communications surveys & tutorials*, vol. 10, no. 4, 2008.

[31] A. Majumda, D. G. Sachs, I. V. Kozintsev, K. Ramchandran, and M. M. Yeung, "Multicast and unicast real-time video streaming over wireless lans," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 524–534, 2002.

[32] S. Ehsan and B. Hamdaoui, "A survey on energy-efficient routing techniques with qos assurances for wireless multimedia sensor networks," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 2, pp. 265–278, 2012.

[33] I. F. Akyildiz, T. Melodia, and K. R. Chowdury, “Wireless multimedia sensor networks: A survey,” *IEEE Wireless Communications*, vol. 14, no. 6, 2007.

[34] B. Hamdaoui and P. Ramanathan, “Energy-efficient and mac-aware routing for data aggregation in sensor networks,” *Proc. of Sensor Network Operations*, 2004.

[35] M. A. Hoque, M. Siekkinen, and J. K. Nurminen, “Energy efficient multimedia streaming to mobile devices—a survey,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 579–597, 2014.

[36] B. Hamdaoui and P. Ramanathan, “Cross-layer optimized conditions for qos support in multi-hop wireless networks with mimo links,” *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, 2007.

[37] T.-W. Chen, J. Tsai, and M. Gerla, “Qos routing performance in multihop, multimedia, wireless networks,” in *Universal Personal Communications Record, 1997. Conference Record., 1997 IEEE 6th International Conference on*, vol. 2. IEEE, 1997, pp. 557–561.

[38] B. Hamdaoui and P. Ramanathan, “Sufficient conditions for flow admission control in wireless ad-hoc networks,” *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 9, no. 4, pp. 15–24, 2005.

[39] F. Soldo, C. Casetti, C.-F. Chiasserini, and P. Chaparro, “Streaming media distribution in VANETs,” in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE.* IEEE, 2008, pp. 1–6.

[40] X. Ge, X. Huang, Y. Wang, M. Chen, Q. Li, T. Han, and C.-X. Wang, “Energy-efficiency optimization for mimo-ofdm mobile multimedia communication systems with qos constraints,” *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2127–2138, 2014.

[41] B. Hamdaoui and P. Ramanathan, “A cross-layer admission control framework for wireless ad-hoc networks using multiple antennas,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 11, 2007.

[42] R. K. Panta, “Mobile video delivery: Challenges and opportunities,” *IEEE Internet Computing*, vol. 19, no. 3, pp. 64–67, May 2015.

[43] K. Piamrat, C. Viho, J. M. Bonnin, and A. Ksentini, “Quality of experience measurements for video streaming over wireless networks,” in *2009 Sixth International Conference on Information Technology: New Generations*, April 2009, pp. 1184–1189.

[44] G. S. Park, W. Kim, S. H. Jeong, and H. Song, "Smart base station-assisted partial-flow device-to-device offloading system for video streaming services," *IEEE Transactions on Mobile Computing*, vol. 16, no. 9, pp. 2639–2655, Sept 2017.

[45] S. Zaidi, S. Bitam, and A. Mellouk, "Enhanced user datagram protocol for video streaming in vanet," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[46] T. A. Q. Pham, K. Piamrat, and C. Viho, "Qoe-aware routing for video streaming over vanets," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Sept 2014, pp. 1–5.

[47] C. Xu, F. Zhao, J. Guan, H. Zhang, and G. M. Muntean, "Qoe-driven user-centric vod services in urban multihomed p2p-based vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 2273–2289, Jun 2013.

[48] E. Baştuğ, M. Bennis, and M. Debbah, *Proactive Caching in 5G Small Cell Networks*. John Wiley & Sons, Ltd, 2016, pp. 78–98. [Online]. Available: http://dx.doi.org/10.1002/9781118979846.ch6

[49] A. Gomes, T. Braun, and E. Monteiro, "Enhanced caching strategies at the edge of lte mobile networks," in *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, May 2016, pp. 341–349.

[50] E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, Aug 2014, pp. 649–653.

[51] W. C. Ao and K. Psounis, "Distributed caching and small cell cooperation for fast content delivery," in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. MobiHoc '15. New York, NY, USA: ACM, 2015, pp. 127–136. [Online]. Available: http://doi.acm.org/10.1145/2746285.2746300

[52] X. Peng, J. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," *CoRR*, vol. abs/1509.00558, 2015. [Online]. Available: http://arxiv.org/abs/1509.00558

[53] E. Baştuğ, J. L. Guénégo, and M. Debbah, "Proactive small cell networks," in *ICT 2013*, May 2013, pp. 1–5.

[54] H. Ahlehagh and S. Dey, "Video caching in radio access network: Impact on delay and capacity," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2012, pp. 2276–2281.

[55] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, April 2013.

[56] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 1107–1115.

[57] D. Liu and C. Yang, "Will caching at base station improve energy efficiency of downlink transmission?" in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 173–177.

[58] Z. Tong, Y. Xu, T. Yang, and B. Hu, "Quality-driven proactive caching of scalable videos over small cell networks," in *2016 12th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, Dec 2016, pp. 90–96.

[59] L. Xiang, D. W. K. Ng, R. Schober, and V. W. S. Wong, "Secure Video Streaming in Heterogeneous Small Cell Networks with Untrusted Cache Helpers," *ArXiv e-prints*, Jul. 2017.

[60] "Linknyc," https://en.wikipedia.org/wiki/LinkNYC.

[61] 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.