Caching and Computing at the Edge for Mobile Augmented Reality and Virtual Reality (AR/VR) in 5G

Melike Erol-Kantarci(✉) and Sukhmani Sukhmani

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada
{melike.erolkantarci,steer031}@uottawa.ca

Abstract. The enormous increase in powerful mobile devices has created hype for mobile data traffic. The demand for high definition images and good quality video streaming for the mobile users has constantly being escalated over the recent decade. In particular, the newly emerging mobile Augmented Reality and Virtual Reality (AR/VR) applications are anticipated to be among the most demanding applications over wireless networks so far. The architecture of the cellular networks has been centralized over the years, which makes the wireless link capacity, bandwidth and backhaul network difficult to cope with the explosive growth in the mobile user traffic. Along with the rise in overall network traffic, mobile users tend to seek similar types of data at different time instants creating a bottleneck in the backhaul link. To overcome such challenges in a network, emerging techniques of caching the popular content and performing computation at the edge are gaining importance. The emergence of such techniques for near future 5G networks would pose less pressure on the backhaul links as well as the cloud servers, thereby, reducing the end-to-end latency of AR/VR applications. This paper surveys the recent edge computing techniques along with the powerful caching strategies at the edge and provides a roadmap for 5G and beyond wireless networks in the context of emerging applications.

Keywords: In-network caching · Mobile edge computing
Mobile augmented reality and virtual reality · Wireless networks · 5G

1 Introduction

The exponential growth in the number of handheld devices such as mobile phones, and tablets has dramatically increased the mobile data traffic. The demand for rich multimedia applications is also rising enormously. This traffic is presumed to steadily increase over the coming years as well. The major source of traffic in the network is due to the demand of video streaming services. Asynchronous content reuse property is exhibited by these live streaming requests of popular content by mobile users that accounts for most of the data traffic. In addition mobile Augmented Reality and Virtual Reality (AR/VR) are expected to be among the first wave of killer applications in 5G. According to ABI Research, the total AR market is expected to reach $114 billion by 2021 while the total VR market is anticipated to reach $65 billion within the same timeframe [1]. AR/VR applications are highly delay sensitive and their performance can degrade

Y. Zhou and T. Kunz (Eds.): AdHocNets 2017, LNICST 223, pp. 169–177, 2018.
https://doi.org/10.1007/978-3-319-74439-1_15

significantly with non-uniform delay and throughput [19]. The future 5G networks are expected to be more fast, flexible, reliable and resilient with round trip time of requests corresponding to 1 ms taking into account the growing mobile traffic. Successful working of technologies like device-to-device communications, millimeter wave and small cell densification can help to achieve the desired parameters for the 5G networks.

Small cell base stations are to be deployed within micro cells, pico cells and femto cells in order to achieve expected to increase capacity and coverage. The deployment of small cell base stations allows resource reuse to a larger extent. Even though placing the small cell base stations in a cell site of various sizes is the most important enabler for higher data rates, the limitation is the backhaul link capacity that provides connection to the core network. These links are most likely wireless due to rapid deployment, self-configuration and cost efficiency [2]. The backhaul links are limited in capacity due to bandwidth constraints and energy consumption while transmitting the packets over long distances.

In order to deal with such issues and better utilization of the limited available resources, the concept of caching has been exploited. Caching of popular contents at the network edge can significantly improve performance as shown in [3, 4]. Delay is reduced along with overcoming backhaul link congestion [2, 4, 5]. By efficiently exploiting the idea of caching at the edge, backhaul cost can be decreased linearly with the base station cache size according to [6, 7] which is possible by the elimination of redundant traffic as also catered by [8]. This further arises the question of what to cache? Popular YouTube videos specially top the list for caching, as they require high data rates for continuous buffering. Then arises the question of where to cache? As caching can be performed at radio access network (RAN) or the Core Network (CN), one has to decide which place would be more beneficial as well as cost efficient. The existing literature on edge-caching focuses on traditional content access, such as watching YouTube videos over mobile devices. However, in AR/VR the access pattern, content to be cached, and the location of the content could be further optimized based on specific applications. For instance, in a museum visit supported by AR, environment maps and constant content can be cached closer to the user while interactive content only might be shared with servers. This can be further extended to caching on D2D networks where communication pattern is more ad hoc.

In case of upcoming 5G networks the feature of antenna directivity [4] can cause retrieval and connection delays for the users with high mobility. Therefore, performing caching at the edge in a distributed fashion can further trigger improved quality of service. There could arise situations where the cached data is not appropriate to provide for the user requests, then the request can forward to one level up in the hierarchy, that is, to the cloud servers. Taking into account another situation where the user is travelling from one cell site to another and some nomadic user is requesting for a particular data, which the first user possesses, and not the BS of that cell site. Then in this case rather than the base station contacting the other BS of different cell site, the user that already has the requested application can share the data locally. This can happen in the context of device-to-device (D2D) communications, which further facilitates faster communication and poses less pressure on the network resources. This exchange of cached content from device-to-device as in [1, 6] allows directed communications between the users.

This however is not currently implemented in practice, as the users are not willing to share any data over D2D links without receiving any rewards in exchange from the network providers or operators.

On the other hand, cloud computing makes it quite convenient to access shared pool of services and resources that are location independent. The idea of cloud computing in the wireless mobile networks leads to the mobile cloud computing and towards cloud radio access networks. For low-latency applications with high computation load, it is not feasible to transmit large amounts of data over long distances to the cloud servers for computation purposes. To address these issues the concept of edge computing or fog computing has emerged recently [9]. Edge computing urges the deployment of computing resources closer to the end users. The edge device could be any device that resides between the data sources and the cloud based data centers. The computing tasks like processing, storing, caching and load balancing can be performed efficiently in edge computing. The benefits of using edge computing can be summarized by saying that, response time is reduced, enhances performance as computations are performed close to the source, network resources are conserved, reduced latency and minimum bottleneck probability. Cloudlets or edge-clouds are considered to be an important part of the 5G network, in particular to support AR/VR applications [10]. It is worth to note that even though, edge computing is a promising technology for achieving better services and higher data rates, they are not capable of replacing the cloud computing. All the heavy applications and intensive computations would likely to be beyond the scope of edge computing at least over the next several years. For decreasing the pressure on the network resources like bandwidth and backhaul links load offloading, in [11] several techniques are discussed for improving the overall efficiency of the network.

In this paper, we summarize the recently proposed techniques in content caching in and edge computing within the wireless networks. We categorize the edge caching techniques according to where the caching is performed; i.e. at the radio access network, core network and the devices Similarly, for edge computing, we group the techniques based on the placement of computing resources either at the cloud or closer to the users. We discuss the impact of the proposed techniques on delay and throughput. Besides, we survey several techniques that focus on energy-efficiency. In closing, we outline the future perspectives and draw a roadmap for new research directions in particular new requirements of mobile AR/VR applications.

The rest of the paper is organized as follows: Sect. 2 discusses the concepts of caching at the edge. Section 3 focuses on computations being performed at the network edge. Section 4 covers the energy-efficiency aspect of proposed techniques. Finally, Sect. 5 conclusions the paper giving future perspectives.

2 Caching at Edge

Videos of major sports events or viral videos over the social networks are accessed by many users at the same time. Bringing content each and every time from the Internet servers creates a bottleneck in the backhaul network. Considering that these users could also be geographically close, caching at the edge has been proposed in the literature. As

the mobility of the user is the main concern for 5G networks, we need appropriate area where caching can be performed. This can be done at the Evolved Packet Core (EPC) or the Radio Access Networks (RAN). We first summarize the techniques that consider caching at EPC and then discuss the studies in RAN.

2.1 Caching Within EPC

EPC includes the serving gateway (S-GW), packet data network gateways (P-GW) and the mobility management entity (MME). The current deployments of cache are done mostly at the P-GW and are known as mobile content delivery networks. In [8], the authors have proposed chunk-level, TPC-level and packet-level caching for EPC. At the chunk-level, the files are initially divided into chunks and then a caching server is used to cache particular chunks. They are then specified and differentiated by hash tags. Further, if the size of the chunk and the hash is the same, the requests related to that chunk are taken care by the same cached chunk. At the TCP-level, TCP flows are managed. Caching intermediates further dividing the file into chunks of fixed or variable length, which helps in scalable cache management. At the packet-level, two middle-boxes at upstream and downstream are used. The role of upstream middle-box is to eliminate redundant bytes whereas the downstream middle-box helps in reconstructing the cached packets. The drawback of this type of caching is that the size of chunks is very small thereby probability of exploding the index sizes in high-speed networks is more.

2.2 Caching at RAN

Caching at RAN is comparatively challenging as tunnels are established between the users and the EPC by the evolved nodeBs (eNB). As the files to be transferred over the connection are first converted into packets and are further encapsulated by GTP tunneling, this make them difficult to perform content aware caching. In order to deal with this situation the concept of byte caching has been implemented. In byte caching, multiple portions of a file are cached at the network layer by searching for the common range of data in the bytes of the packet flows. It does not subdivide the packet flows into fragments, but aim at caching the frequently used bytes in the flows, thereby deleting the redundant ones.

As the caching memory of the base stations or eNBs is limited, complex caching schemes are required to be followed in order to gain more flexibility and cost effective network. There should also be some collaborative schemes installed in the neighboring eNBs for achieving successful RAN caching output. Moving further towards the distributed caching of videos at the base station of RAN, can increase the efficiency of the delivering the content even more. According to [12] caching at RAN with the help of User Preference Profile (UPP) along with the video aware backhaul scheduling, can increase the capacity significantly when compared to the traditional techniques. Considering the scenario in [4] the streaming of videos can be achieved with low handoff delays and less connection latency using caching at the RAN. The streaming of live data or any video also accounts for the base station cache size according to [6]. Their main focus is

reducing the transmit power cost, with increased base station cache, which further results in linear decrement of the overall backhaul cost.

2.3 Device-to-Device (D2D) Caching

In [1], the authors consider caching at the small base stations as well as the user terminals, which can carry out their communication using D2D communications. Working of transmission and caching protocols together yield two types of gains that can be evaluated according to [1]. First being the local caching gain, this is achieved when any device withholding pre-cached information about the data locally satisfies the content requested by the user. Second, global caching gain can be achieved. This is a cost effective way of caching for the service providers. The general assumption here is that the user demands are known in advance. When caching is performed at the user terminal, energy consumption over backhaul networks can be also reduced [13, 14]. D2D caching can also borrow some concepts from ad hoc networking which has not been explored so far.

2.4 Transcoding Enabled Caching (TeC)

According to [15], another way of increasing the quality of the video output as requested dynamically by the mobile users is, by using the transcoding enabled caching technique. The combination of transcoding and caching can serve the heterogeneous users in two of the efficient ways: first, by decreasing the user estimated latency; second, by reducing the traffic between the proxy and the main server. This technique works as follows. Two types of transcoders, namely bit-rate reduction and spatial resolution reduction transcoding are introduced. Transcoding unit is placed on the content delivery path such that depending on the connection speed and processing capability of an end user, the content is converted into an appropriate format.

3 Mobile Computing

With the skyrocketing use of smart devices on-the-go, computing on mobile devices, i.e. mobile computing, became an essential part of devices. The limited resources of mobile devices require computational help either from cloud servers or other resources around them, in particular for performing computationally heavy tasks. Mobile Edge Computing (MEC) refers to performing such computations on locations closer to the user than the cloud servers. The very concept of cloud is to provide software that is capable of executing intensive and heavy computations. They are convenient to use and caters the dynamic requests by accessing a shared pool of various configurable devices [16]. Cloud servers can compute extensive applications efficiently, thereby, increasing the battery life of the mobile devices [17]. If the applications that require heavy computation are shared to the smaller version of clouds in the vicinity of the device these are referred to as cloudlets which is a key concept for MEC [18].

3.1 Computing at the Edge

Following the newly emerging AR/VR applications and many other delay-sensitive applications in 5G, it is apparent that providing the required low-latency with cloud computing and the transfer of massive amounts of data to the data centers may not be possible or could be not economical [16]. In order to overcome these drawbacks the concept of edge computing emerges as a promising tool [18]. Edge computing also referred as fog computing, comprises of proxy servers that are located at the network edge. The very idea of edge computing is to achieve low latency and location awareness. It is worth to note that edge computing is not capable of replacing cloud, as heavy applications cannot be realized through them but they can reduce the burden of the cloud servers by locally serving the requests generated by the mobile users. As the cloud and cloudlets are expected work in harmony, their interoperability emerges as a research focus. In the next section, we first focus on the techniques that study this challenge.

3.2 Cloud-Fog Interoperability

In [20], the authors suggest using software defined networking architecture to integrate the infrastructure of cloud and fog. According to the architecture, interoperability can be of two ways:

- Fog-Fog Computing: When the data requested by the mobile user is beyond the scope of a single fog server, it can seek that data from another nearby fog server. This is referred to as fog-fog computing.
- Fog-Cloud Computing: When the service being requested by the user is beyond the scope of the fog server, then the request is forwarded to the cloud server in order to provide the user with the required service.

In addition, caching when combined with MEC can increase the quality of service significantly. In the next section, we summarize the works, that focus on the energy-efficiency improvement when these two approaches are combined.

4 Energy Efficient Caching and Computing

Various techniques of offloading and load balancing are required to satisfy diverse types of requests with balanced energy consumption. For instance, the dynamic offloading framework in [21] helps to balance the load efficiently in a network. Cloud computing differs from traditional models due to the adoption of virtualization. This very feature of cloud leverages the operators to run arbitrary applications from various customers on the virtual machines. The cloud providers reduce the energy consumption on the mobile systems by providing computing cycles to the users aiming to decrease the computation at the mobile user end.

Mobile devices are equipped with multiple network interfaces. The server interface comprises of EDGE, UMTS and GPRS, which are responsible for corresponding with the network operators. The other is the peer interface that includes Bluetooth and IEEE 802.11 that connects to other computing devices [22]. The joint use of these interfaces

can lead to energy efficient data applications according to [22]. In addition, as evidenced by [23], most of the studies do not consider energy consumption at the backhaul, however it also contributes a significant amount of energy consumption. Therefore, the potential of MEC to enhance energy-efficiency is supported by their reduced load at the backhaul links.

Moving further towards the caching domain, a generalized notion of saving energy is to simply turn a device down. However, for content-caching with D2D, this inflicts inconsistency in the already cached content. In that case, once caching is performed cache invalidation strategy needs to be implemented to make sure that the data cached in the mobile device is copied to the server. According to [24] cache consistency is difficult to enforce when the mobile devices are powered off.

Procuring the energy efficient mechanisms in the 5G networks, the content placement issue is addressed in [25]. The proposed framework is CAR (cache-at-relay) comprises of three integer linear programming models that aims to reduce device power consumption by uplink power optimization. The impact of caching on the backhaul links is discussed in [26] where several techniques are shown to relax the load on the backhaul. In mobile AR/VR applications battery is one of the major bottlenecks. Therefore energy-efficient techniques are expected to play a key role when it comes to adoption of edge-caching or edge-computing techniques.

5 Conclusions and Future Directions

5G and beyond networks are aiming to serve many applications that desire ultra-low latency. For instance, mobile AR/VR applications, tactile internet and autonomous driving require latency values close to 1 ms. They also require frequent video or map access and streaming content which calls for placing caching and computing resources closer to mobile users. This paper surveys the recent studies that focus on caching and computing at the edge. Various types of caching techniques have been discussed at the network core as well as the network edge. The combination of various kinds of caching techniques that reduce the duplicity and eliminate redundant traffic in the network have been summarized. Further, mobile edge computing techniques have been discussed. Computations can be done either at the cloud server or locally at the edge server. Though the cloud-based servers are designed to support heavy and complex computations, the fog servers are capable of catering less complex applications. They can be implemented at the edge, which would further take into account the often-requested content and their output.

The major focus of research in the literature has been enhancing the quality of service and energy-efficiency. Although these are relevant to AR/VR and other ultra-low latency applications, next-generation wireless networks will need more flexibility and configurability capabilities. With this regard, Software Defined Networking (SDN) is expected to play a critical role in increasing the flexibility of caching and computing in mobile networks. SDN allows heterogeneous devices to be programmed by the controller dynamically. The proper coordination of the controller management along with the device executions, caching and computations can enhance AR/VR experience.

Furthermore, it is apparent that 5G and beyond wireless networks will have significant differences than today's LTE networks. The high amount of investments suggests that the architecture needs to be flexible enough to support various applications. In other words, functionality to support AR/VR applications should, for example, support autonomous car applications. As a result, application-specific platform services need to be dynamically adapted. Caching and computing at the edge is a powerful tool to support dynamicity however scalable control of large number of distributed systems becomes the challenging part. In addition, cross-application performance uniformity will emerge as a significant challenge. As a future direction, advanced algorithms implemented over SDN are expected to play an important role in reconfiguring 5G networks to satisfy the demands of newly emerging applications.

References

1. ABI Research and Qualcomm: Augmented and Virtual Reality: the First Wave of 5G Killer Apps. White paper (2017). https://www.qualcomm.com/news/onq/2017/02/01/vr-and-ar-are-pushing-limits-connectivity-5g-our-rescue
2. Gregori, M., Gómez-Vilardebó, J., Matamoros, J., Gündüz, D.: Wireless content caching for small cell and D2D networks. IEEE J. Sel. Areas Commun. **34**(5), 1222–1234 (2016)
3. Goebbels, S., Jennen, R.: Enhancements in wireless broadband networks using smart caching an analytical evaluation. In: Proceedings of International Symposium Personal, Indoor and Mobile Radio Communications, pp. 1–5, September 2008
4. Qiao, J., He, Y., Shen, X.S.: Proactive caching for mobile video streaming in millimeter wave 5G networks. IEEE Trans. Wireless Commun. **15**(10), 7187–7198 (2016)
5. Dehghan, M. et al.: On the complexity of optimal routing and content caching in heterogeneous networks. In: Proceedings of IEEE Conference on Computer Communications, pp. 936–944, April 2015
6. Liu, A., Lau, V.K.N.: Exploiting base station caching in MIMO cellular networks: opportunistic cooperation for video streaming. IEEE Trans. Signal Process. **63**(1), 57–69 (2015)
7. Golrezaei, N., Molisch, A.F., Dimakis, A.G., Caire, G.: Femtocaching and device-to-device collaboration: a new architecture for wireless video distribution. IEEE Commun. Mag. **51**(4), 142–149 (2013)
8. Wang, X., Chen, M., Taleb, T., Ksentini, A., Leung, V.C.M.: Cache in the air: exploiting content caching and delivery techniques for 5G systems. IEEE Commun. Mag. **52**(2), 131–139 (2014)
9. Satyanarayanan, M.: The emergence of edge computing. Computer **50**(1), 30–39 (2017)
10. Seam, A., Poll, A., Wright, R., Mueller, J., Hoodbhoy, F.: Enabling mobile augmented and virtual reality with 5G networks. ATT White paper (2017). http://about.att.com/content/dam/innovationblogdocs/Enabling%20Mobile%20Augmented%20and%20Virtual%20Reality%20with%205G%20Networks.pdf
11. Kumar, K., Lu, Y.H.: Cloud computing for mobile users: can offloading computation save energy? Computer **43**(4), 51–56 (2010)
12. Ahlehagh, H., Dey, S.: Video-aware scheduling and caching in the radio access network. IEEE/ACM Trans. Networking **22**(5), 1444–1462 (2014)
13. Ji, M., Caire, G., Molisch, A.F.: Wireless device-to-device caching networks: basic principles and system performance. IEEE J. Sel. Areas Commun. **34**(1), 176–189 (2016)

14. Ji, M., Caire, G., Molisch, A.F.: Fundamental limits of caching in wireless D2D networks. IEEE Trans. Inf. Theory **62**(2), 849–869 (2016)
15. Shen, B., Lee, S.-J., Basu, S.: Caching strategies in transcoding-enabled proxy systems for streaming media distribution networks. IEEE Trans. Multimedia **6**(2), 375–386 (2004)
16. Huo, R., Yu, R., Huang, T., Xie, R., Liu, J., Leung, V.C.M., Liu, Y.: Software defined networking, caching, and computing for green wireless networks. IEEE Commun. Mag. **54**(11), 185–193 (2016)
17. Goudarzi, M., Movahedi, Z., Nazari, M.: Mobile cloud computing: a multisite computation offloading. In: 8th International Symposium on Telecommunications (IST), Tehran, pp. 660–665 (2016)
18. Satyanarayanan, M., Bahl, P., Caceres, R., Davies, N.: The case for VM-based cloudlets in mobile computing. IEEE Pervasive Comput. **8**(4), 14–23 (2009)
19. Westpal, C.: Challenges in networking to support augmented reality and virtual reality. In: ICNC 2017 (2017)
20. Yang, P., Zhang, N., Bi, Y., Yu, L., Shen, X.: Catalyzing cloud-fog interoperation in 5G wireless networks: an SDN approach. Technical report, April 2017. https://arxiv.org/pdf/1612.05291.pdf
21. Kosta, S., Aucinas, A., Hui, P., Mortier, R., Zhang, X.: ThinkAir: dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In: Proceedings IEEE INFOCOM, Orlando, FL, pp. 945–953 (2012)
22. Yeung, M.K.H., Kwok, Y.-K.: A game theoretic approach to energy efficient cooperative cache maintenance in MANETs. In: IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, Berlin, vol. 3, pp. 1500–1504 (2005)
23. Huq, K.M.S., Mumtaz, S., Bachmatiuk, J., Rodriguez, J., Wang, X., Aguiar, R.L.: Green HetNet CoMP: energy efficiency analysis and optimization. IEEE Trans. Veh. Technol. **64**(10), 4670–4683 (2015)
24. Wu, K.-L., Yu, P.S., Chen, M.-S.: Energy-efficient caching for wireless mobile computing. In: Proceedings of the Twelfth International Conference on Data Engineering, New Orleans, LA, pp. 336–343 (1996)
25. Erol-Kantarci, M.: Cache-at-relay: energy-efficient content placement for next-generation wireless relays. Int. J. Netw. Manage. **25**, 454–470 (2015)
26. Bahmani, K., Argyriou, A., Erol-Kantarci, M.: Backhaul relaxation through caching. In: Imran, M., Raza, S.A., Shakir, M.Z. (eds.) Access, Fronthaul and Backhaul for 5G Wireless Networks. IET (2017)