Latency-aware Cache Policy in Multi-Access Edge Computing: A Deep Reinforcement Learning Approach

Marisangila Alves
Graduate Program in Applied Computing - Santa Catarina State University - UDESC
marisangila.alves@udesc.br

I. Introduction

Mobile devices are largely present in daily activities and have become the most used form of Internet access for end-users. In this context, an evolution in mobile networks is happening to support new applications and services. The Fifth Generation Technology Standard (5G) networks are being implemented and are becoming effectively available in some countries posing new management and administrative challenges to Mobile Network Operators (MNOs) [1]. Among the major Quality-of-Service (QoS) requirements defined by 5G, the ultra low latency and the high throughput between end-users and cloud or edge based services deserve to be highlighted. The former opens the opportunity to popularize applications as virtual and augmented reality, Industry 4.0, autonomous vehicles, among others, which have a strict latency requirement [2], while the latter is required by applications based on data transfer operations and mobile video traffic [3].

The physical and logical proximity between resources (services, storage and computing) and end-users in Heterogeneous Cellular Network (HCN) is essential to deliver the 5G QoS requirements [4], [5]. Specifically, the placement of caches on Radio Access Network (RAN), Multi-Access Edge Computing (MEC), and low-power nodes are natural choice to decrease end-to-end latency, increase application's throughput, and to reduce the replicated content load in backhaul network [6], [7]. In this scenario, we claim that the cache placement and requests routing must be jointly performed to deliver QoS for new and running applications. First and foremost, the mobility of end-users on the RAN infrastructures poses a challenge regarding the dynamic routing of data between mobile devices, cache replicas eventually placed on Base Stations (BSs), and external repositories accessed through the backhaul network. Secondly, the heterogeneity of applications requires distinct cache configurations to host multiple concurring users (e.g. a data-sharing application requires more storage cache, while a web page server may require more memory).

Although the specialized literature largely focused on developing caching policies to cache placement [8] and data routing approaches to improve the QoS [9]–[14], the existing approaches do not consider cooperation [8], [9], [15]. Some approaches consider the cooperation only BSs neighbors (one-hop) [10], [11] or decrease the search for content in RAN through hierarchical cooperation [12], [13]. There are strategies which consider multi-hops request routing and cooperation; Nonetheless, the mobility is not considered [14]. Indeed, some proposals ignore the fact that the network can be used by many applications, not just for delivering the services managed by the cache system.

In this sense, this work proposes a cooperative policy aiming the joint placement of caches and users' requests routing on HCNs, which objective is to minimize the latency.

The policy innovates by applying well-known TCP fundamentals to infer formation about the network infrastructure at application layer, specifically bandwidth and Round-Trip Time (RTT) values. By combining networking and cache QoS requirements, the policy balances the network load (to help avoiding network congestion) and performs a dynamic cache to HCN resource mapping considering the actual link capacity, instead of only analyzing the maximum link bandwidth.

This problem is based on the multi-commodity flow problem, and the formulations in exact approach methods require high computing power, driven by the combinatorial nature of problem. Other works in the literature have implemented Deep Reinforcement Learning (DRL), which has shown to be a promising approach. Unlike Deep Neural Networks (DNN), DRL can explore the context of users and networks to optimize cache policies [16]. The goal of this study is to compare the optimal solution obtained by the model formulated as a Linear Programming (LP) model presented by the authors in [17] with a model that uses DRL techniques.

References

[1] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Comm. Surveys and Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5g wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[3] I. T. U. ITU, "Minimum requirements related to technical performance for imt-2020 radio interface(s)," Tech. Rep., 2017.

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

[5] 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, 2013.

[6] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. LE, L. B. LE, W.-J. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020.

[7] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[8] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, 2013.

[9] M. Dehghan, B. Jiang, A. Seetharam, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the complexity of optimal request routing and content caching in heterogeneous cache networks," *IEEE/ACM Trans. on Networking*, vol. 25, no. 3, pp. 1635–1648, 2017.

[10] L. Pu, L. Jiao, X. Chen, L. Wang, Q. Xie, and J. Xu, "Online resource allocation, content placement and request routing for cost-efficient edge caching in cloud radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 8, pp. 1751–1767, 2018.

[11] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382–1393, 2017.

[12] X. Li, X. Wang, K. Li, Z. Han, and V. C. M. Leung, "Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design," *IEEE Trans. on Wireless Comm.*, vol. 16, pp. 6926–6939, 2017.

[13] M. Sheng, C. Xu, J. Liu, J. Song, X. Ma, and J. Li, "Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 70–76, 2016.

[14] Y. Song, T. Wo, R. Yang, Q. Shen, and J. Xu, "Joint optimization of cache placement and request routing in unreliable networks," *Journal of Parallel and Distributed Computing*, vol. 157, pp. 168–178, 2021.

[15] D. Harutyunyan, A. Bradai, and R. Riggio, "Trade-offs in cache-enabled mobile networks," in *2018 14th International Conference on Network and Service Management (CNSM)*, 2018, pp. 116–124.

[16] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 869–904, 2020.

[17] M. Alves and G. P. Koslovski, "Network-aware cache provisioning and request routing in heterogeneous cellular networks," *International Journal of Communication Networks and Distributed Systems*, 2024.