On the Performance Gains of Decentralized Edge Caching Schemes in 5G and Beyond

Lilian C. Mutalemwa and Seokjoo Shin*
Department of Computer Engineering, Chosun University, Gwangju 61452, South Korea
Email: lilian.mutalemwa@gmail.com, *sjshin@chosun.ac.kr (corresponding author)

***Abstract*—Edge caching is one of the key technologies that enable low-latency communications in 5G and beyond 5G networks. Performance of an edge caching scheme may vary depending on the type of learning model. Thus, edge caching schemes present different performance when centralized or decentralized learning algorithms are used. In this study, we outline the performance gains of decentralized edge caching. We consider various edge caching frameworks based on deep learning (DL), deep reinforcement learning (DRL), and federated learning (FL) algorithms. It is shown that decentralized frameworks present viable mechanisms to outperform centralized frameworks, especially in dynamic and heterogeneous large-scale networks or complex IoT environments.**

Keywords—5G; ultra-reliable low-latency communications; edge caching; decentralized edge caching; federated learning.

I. INTRODUCTION

Ultra-reliable low-latency communications (URLLC) is probably the most talked-about 5G and beyond 5G (B5G) use case mainly because of the huge services it can support [1]. URLLC aims to deliver a vastly reliable mobile wireless network with extremely low latency requirements. Edge caching is one of the key technologies that facilitate low-latency communications in 5G and B5G networks. Particularly, mobile edge caching (MEC) has been regarded as a promising technique to provide low latency for content access [1]-[3]. In the MEC systems, popular contents can be cached in proximity to the edges of networks in edge devices (EDs), e.g. base stations (BSs) and user equipment (UE) (or mobile devices), which reduces massive duplicated traffic of content deliveries via backhaul networks and shortens the content delivery latency [3]-[26]. An example use case of a MEC system may be video caching on the network edge. At present, video delivery is the dominant traffic in the network. In the year 2022, more than 82% of network traffic will be video traffic, and this number will increase to 90% by the year 2025 [5]. In video delivery, reducing network delay is one of the key factors for improving users' quality of experience (e.g., significantly reduce the playback latency). Therefore, video caching is pushed from the cloud server to the edge network to provide end-users a low-delay video delivery [5].

Employing an appropriate caching algorithm is pivotal to increase the overall quality of experience in content distribution systems as 1% increase in hit rate can have a positive impact [1]. Therefore, in this study, we investigate the performance of centralized and decentralized edge caching solutions. We consider various edge caching frameworks based on deep learning (DL), deep reinforcement learning (DRL), and federated learning (FL) algorithms. Subsequently, we highlight the performance gains of the decentralized edge caching frameworks.

II. CENTRALIZED AND DECENTRALIZED EDGE CACHING SCHEMES

In general, edge caching-enabled systems may be identified as centralized or decentralized, based on their learning models [7], [27]. Due to bandwidth, storage, and privacy concerns, centralized edge caching systems are often impractical [26]. The centralized caching algorithms may result in overconsumed network resources during the training and data transmission processes [9]. The effects of centralized caching schemes can become severe in dynamic and heterogeneous large-scale systems. The centralized edge caching may be realized through DL or DRL algorithms. To address the challenges of centralized edge caching, decentralized caching schemes are considered. The decentralized edge caching frameworks may be realized through DL, DRL, or FL algorithms. Thus, all FL schemes are decentralized. FL is becoming increasingly popular in 5G and B5G networks because it is effective and privacy-preserving by design [28], [29].

Decentralized edge caching schemes are often considered because it is assumed that most 5G and B5G networks will be based on decentralized and infrastructureless communication to enable devices to cooperate directly over device-to-device spontaneous connections [30].

III. PERFORMANCE INVESTIGATIONS

In our investigations, we consider various edge caching solutions based on DL, DRL, and FL learning algorithms. Also, multi-agent DRL (MADRL) and distributed DL (DDL) are considered. The solutions are adopted from [7], [9], [17], [19]-[21], [31], and [32]. Then, we examine the performance of the solutions. Our observations are based on the analysis and evaluations presented in [7], [9], [17], [19]-[21], [31], [32]. A summary of the observations is presented in Table I.

The performance of the DL, DRL, and FL edge caching solutions is compared to the performance of two baseline solutions, the least recently used (LRU) and least frequently used (LFU) caching solutions which are commonly used by content providers [7], [21]. In the LRU algorithm, the system keeps track of the most recent requests for every cached data

TABLE I. PERFORMANCE FEATURES OF CENTRALIZED AND DECENTRALIZED EDGE CACHING SCHEMES.

Technique	Limitations	Example solution	Objective	Learning model	Performance
DL-based edge caching	• Traditional DL-based optimization and prediction schemes take a long running time of recursions for converging to the optimal model. • In centralized schemes, sending streams of raw training data to server can increase network traffic and energy consumption. • Most schemes cannot handle non-IID data or privacy preservation issues.	DLs [19]	• Reduce latency and backhaul network traffic for 5G mobile video streaming.	• Decentralized DL algorithm.	• Lower latency than LRU and LFU. • Higher cache hit rate than LRU and LFU.
		DLs2 [17]	• Reduce service delay for UEs and error of content demand prediction.	• Centralized DL algorithm.	• Lower latency than LRU and LFU. • Higher cache hit rate than LRU and LFU. • Higher latency than DDLs. • Lower cache hit rate than DDLs.
		DDLs [17]	• Reduce service delay for UEs and error of content request prediction while preserving privacy of UEs data.	• Decentralized DDL algorithm.	• Lower latency than LRU, LFU and DLs2. • Higher cache hit rate than LRU, LFU and DLs2.
DRL-based edge caching	• Requires intensive computation capacity for finding optimal model particularly in large-scale data. • Achieves reduced performance when UEs and network states are heterogeneous. • In large-scale data with massive UEs, centralized DRL schemes incur increased traffic on uplink wireless channels. • In large-scale data with massive UEs, it is challenging to perform decentralized DRL due to relatively weak computation capability of UEs. Also, it takes long time to train the DRL agent. • Decentralized DRL increases the energy cost at the UEs. • Most schemes cannot handle unbalanced and non-IID data or privacy preservation issues.	DRLs [32]	• Minimize communication cost and loss of data freshness in IoT applications.	• Centralized DRL algorithm.	• Significantly lower latency than LRU and LFU. • Significantly higher cache hit rate than LRU and LFU. • Higher latency than MADRLs. • Lower cache hit rate than MADRLs.
		MADRLs [21]	• Minimize content access latency and traffic cost in diversified 5G video streaming environment.	• Decentralized MADRL algorithm.	• Significantly lower latency than LRU and LFU. • Significantly higher cache hit rate than LRU and LFU. • Lower latency than DRLs. • Higher cache hit rate than DRLs.
		DRLs2 [7]	• Improve cache hit rate in media-enabled applications.	• Centralized DRL algorithm.	• Higher cache hit rate than LRU and LFU but lower than MADRLs2.
		MADRLs2 [7]	• Reduce latency and improve cache hit rate in media-enabled applications.	• Decentralized MADRL algorithm.	• Lower latency than LRU, LFU, and DRLs2. • Higher cache hit rate than LRU, LFU, and DRLs2.
FL-based edge caching	• When traditional algorithms such as FedAvg is used, FL suffers from a large number of communication rounds to convergence with non-IID datasets. Also, has high communication overhead.	FedDRLs [9]	• Reduce latency, performance loss, and backhaul traffic while improving hit rate in IoT systems.	• FL-based DRL algorithm.	• Significantly lower latency than LRU and LFU. • Significantly higher cache hit rate than LRU and LFU. • Latency and cache hit rate are comparable to a centralized DRL scheme.
		FedDRLs2 [20]	• Make mobile communication system cognitive and adaptive, reduce network traffic, and achieve near-optimal performance with low overhead of learning.	• FL-based DRL algorithm.	• Significantly higher cache hit rate than LRU and LFU. • Cache hit rate is comparable to a centralized DRL scheme.
		FedDRLs3 [31]	• Reduce transmission costs between IoT devices and EDs.	• FL-based DRL algorithm.	• Transmission time is comparable to a centralized DRL.

content. When the cache storage becomes full, the cached content which is requested least recently, is replaced by the new content [7], [9]. For the LFU algorithm, the system keeps track of the number of requests for every cached content. When the cache storage becomes full, the cached content which is least frequently requested, is replaced by the new content [7], [9]. We compare the performance of the solutions in terms of content delivery latency and cache hit rate. The cache hit rate is used to show how frequently the requested content is found in the local cache [7], [21]. Therefore, we assume that cache hit rate indicates the content acquisition reliability. That means, a high cache hit rate corresponds to high content acquisition reliability.

It is shown in Table I that the DL, DRL, and FL edge caching solutions are capable of achieving improved performance in terms of content delivery latency and cache hit rate to outperform the LRU and LFU solutions. The main reason for the poor performance in the LRU and LFU solutions is that, the algorithms in LRU and LFU do not consider the popularity of contents in the future. As a result, the solutions do not adapt well to the dynamically changing content popularity and they achieve low cache efficiency [24], [33]. For instance, the LFU framework is not able to reach a good performance in IoT environment because it does not consider the saltation and timeliness of the IoT data popularity [33]. It was also shown in [9], [19], [21] that the DL, DRL, and FL edge caching solutions are capable of achieving reduced backhaul network traffic to outperform the LRU and LFU solutions.

On the other hand, Table I shows that although the solutions with centralized DL, DRL, and FL algorithms perform better than the LRU and LFU solutions, the centralized solutions present reduced performance when compared to the solutions with decentralized algorithms. For example, it was shown in [17] that a DDL solution performs better than a centralized DL solution in terms of latency and cache hit rate. Furthermore, since the DDL framework only needs to collect the trained models from the EDs without considering any raw dataset transmission, the DDL framework was able to achieve reduced communication overhead. Also, the DDL was able to learn the dataset faster than the centralized DL as the number of the EDs was increased. It was also presented in [7], [21] that decentralized MADRL frameworks achieve better performance than centralized DRL frameworks. As an example, when a massively vibrant, diversified, and distributed video streaming environment was considered in [21], a decentralized MADRL framework presented better performance than a centralized DRL framework in terms of video access latency and the traffic cost. In [20], it was revealed that a FL scheme can consume significantly lower communication resources than a centralized DRL framework. In [20], [31], it was presented that the FL-based solutions achieve a lower number of dropped tasks, queuing delay, and transmission energy to outperform the DRL solutions.

Conversely, when the performance of the FL frameworks was investigated in [9], [20], [31], it was revealed that although FL schemes can address the challenges of DRL and DL frameworks, the FL schemes are less capable of achieving significantly improved performance in terms of content access latency and cache hit rate. For example, the performance of the FL schemes in [9], [20], [31] was comparable to the performance of the centralized DRL schemes once the model aggregation of FL was performed several times. That means, for the FL algorithms to achieve the performance level of the centralized DRL algorithms in terms of content access latency and cache hit rate, the FL algorithms must allow several rounds of model aggregation.

On the other hand, it was pointed out in [9], [31] that the performance level of the FL schemes is reasonable since the FL schemes assume more practical network conditions. As an example, the centralized DRL scheme considered in [9], [20], [31] assumed that the massive training data can be successfully uploaded to the ED without loss or delay. Considering the limitations of wireless channels, the assumption made by the DRL scheme may be impractical. Moreover, the work in [20], [34] considered the challenge that when not independent and identically distributed (non-IID) datasets are used, FL algorithms incur large number of communication rounds to converge to the global optimal. Consequently, [20], [25], [34] highlighted the technique of transfer learning as a potential solution for the challenge. Thus, [20], [25], [34] considered the use of transfer learning technique to improve the learning efficiency of the FL algorithms. It was pointed out in [34] that the transfer learning technique can ensure training is not initialized from scratch. In [25], it was demonstrated that personalized federated learning can significantly reduce the performance degradations caused by the non-IID data.

Several other benefits of using FL edge caching were highlighted in [15], [20], [22], [27], [34]. The following benefits were outlined: the system becomes more cognitive and robust, improved flexibility, reduced network traffic and energy consumption, privacy preservation, and improved stability despite loss of connectivity.

IV. CONCLUSION AND FUTURE WORK

This paper presents some investigations on the performance features of centralized and decentralized edge caching schemes. It is shown that decentralized frameworks present better performance than centralized frameworks in terms of content delivery latency and traffic cost. Moreover, the decentralized frameworks perform better in a massively vibrant, diversified, and distributed video streaming environment or in dynamic and heterogeneous large-scale networks where devices are resource-constrained, including in complex IoT environments. Furthermore, it is shown that FL edge caching presents viable mechanisms in 5G and B5G networks. However, traditional FL algorithms incur high communication overhead. Therefore, as part of our future work, we will explore the techniques to improve the communication efficiency of FL algorithms. In particular, we

will study the mechanisms of transfer learning and personalized FL.

ACKNOWLEDGMENT

This research is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07048338).

REFERENCES

[1] E. E. Ugwuanyi, M. Iqbal and T. Dagiuklas, "A novel predictive-collaborative-replacement (pcr) intelligent caching scheme for multi-access edge computing," *IEEE Access*, vol. 9, pp. 37103–37115, March 2021.

[2] J. Liang *et al*., "Multi-head attention based popularity prediction caching in social content-centric networking with mobile edge computing," *IEEE Communications Letters*, vol. 25, no. 2, pp. 508–512, Feb. 2021.

[3] X. Wang, R. Li, C. Wang, X. Li, T. Taleb and V. C. M. Leung, "Attention-weighted federated deep reinforcement learning for device-to-device assisted heterogeneous collaborative edge caching," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 154–169, Jan. 2021.

[4] X. Xia, F. Chen, Q. He, J. Grundy, M. Abdelrazek and H. Jin, "Online collaborative data caching in edge computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 2, pp. 281–294, Feb. 2021.

[5] Y. Guan, X. Zhang and Z. Guo, "PrefCache: Edge cache admission with user preference learning for video content distribution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1618–1631, April 2021.

[6] S. K. Sharma, I. Woungang, A. Anpalagan, and S. Chatzinotas, "Toward tactile Internet in beyond 5G era: Recent advances, current issues, and future directions," *IEEE Access*, vol. 8, pp. 56948–56991, 2020.

[7] C. Zhong, M. C. Gursoy, and S. Velipasalar, "Deep reinforcement learning-based edge caching in wireless networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 1, pp. 48–61, Mar. 2020.

[8] M. Mehrabi, D. You, V. Latzko, H. Salah, M. Reisslein, and F. H. P. Fitzek, "Device-enhanced MEC: Multi-access edge computing (MEC) aided by end device computation and caching: A survey," *IEEE Access*, vol. 7, pp. 166079–166108, Nov. 2019.

[9] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, "Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9441–9455, Oct. 2020.

[10] S. He, J. Ren, J. Wang, Y. Huang, Y. Zhang, W. Zhuang, and S. Shen, "Cloud-edge coordinated processing: Low-latency multicasting transmission," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1144–1158, May 2019.

[11] T. Zhang, X. Fang, Y. Liu, and A. Nallanathan, "Content-centric mobile edge caching," *IEEE Access*, vol. 8, pp. 11722–11731, Jan. 2020.

[12] U. Paul, J. Liu, S. Troia, O. Falowo, and G. Maier, "Traffic-profile and machine learning based regional data center design and operation for 5G network," *J. Commun. Netw.*, vol. 21, no. 6, pp. 569–583, Dec. 2019.

[13] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 8, pp. 1791–1805, Aug. 2018.

[14] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for lowlatency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

[15] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Federated learning for ultra-reliable low-latency V2 V communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.

[16] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1610.02527

[17] Y. M. Saputra *et al.*, "Distributed deep learning at the edge: A novel proactive and cooperative caching framework for mobile edge networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1220–1223, Aug. 2019.

[18] X. Fan, Y. Huang, X. Ma, J. Liu, and V. C. M. Leung, "Exploiting the edge power: An edge deep learning framework," *CCF Trans. Netw.*, vol. 2, no. 1, pp. 4–11, Dec. 2018.

[19] H. Pang, J. Liu, X. Fan, and L. Sun, "Toward smart and cooperative edge caching for 5G networks: A deep learning based approach," in *Proc. IWQoS*, Jun. 2018, pp. 1–6.

[20] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Sep. 2019.

[21] F. Wang, F. Wang, J. Liu, R. Shea, and L. Sun, "Intelligent video caching at network edge: A multi-agent deep reinforcement learning approach," in *Proc. IEEE INFOCOM*, Jul. 2020, pp. 2499–2508.

[22] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities and challenges," 2019, *arXiv:1908.06847*. [Online]. Available: http://arxiv.org/abs/1908.06847

[23] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Differentially private asynchronous federated learning for mobile edge computing in urban informatics," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2134–2143, Mar. 2020.

[24] Z. Yu, J. Hu, G. Min, H. Lu, Z. Zhao, H. Wang, and N. Georgalas, "Federated learning based proactive content caching in edge computing," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[25] Q. Wu, K. He, and X. Chen, "Personalized federated learning for intelligent IoT applications: A cloud-edge based framework," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 35–44, May 2020.

[26] S. Wang *et al.*, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

[27] L. C. Mutalemwa and S. Shin, "A classification of the enabling techniques for low latency and reliable communications in 5g and beyond: Ai-enabled edge caching," *IEEE Access*, vol. 8, pp. 205502-205533, Nov. 2020.

[28] S. Savazzi, M. Nicoli, M. Bennis, S. Kianoush and L. Barbieri, "Opportunities of federated learning in connected, cooperative, and automated industrial systems," *IEEE Communications Magazine*, vol. 59, no. 2, pp. 16-21, Feb. 2021.

[29] J. Posner, L. Tseng, M. Aloqaily and Y. Jararweh, "Federated learning in vehicular networks: Opportunities and solutions," *IEEE Network*, vol. 35, no. 2, pp. 152-159, April 2021.

[30] S. Savazzi, M. Nicoli, and V. Rampa, "Federated learning with cooperating devices: A consensus approach for massive IoT networks," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4641–4654, May 2020.

[31] J. Ren, H. Wang, T. Hou, S. Zheng, and C. Tang, "Federated learning-based computation offloading optimization in edge computing-supported Internet of Things," *IEEE Access*, vol. 7, pp. 69194–69201, Jun. 2019.

[32] H. Zhu, Y. Cao, X. Wei, W. Wang, T. Jiang, and S. Jin, "Caching transient data for Internet of Things: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2074–2083, Apr. 2019.

[33] B. Chen, L. Liu, M. Sun, and H. Ma, "IoTCache: Toward data-driven network caching for Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10064–10076, Dec. 2019.

[34] W. Y. B. Lim *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 3rd Quart., 2020.