Learning-based Content Delivery in 5G-enabled Multi-access Edge Computing

Erfan Farhangi Maleki
University of Delaware
Newark, DE, USA
erfanf@udel.edu

Weibin Ma
University of Delaware
Newark, DE, USA
weibinma@udel.edu

Lena Mashayekhy
University of Delaware
Newark, DE, USA
mlena@udel.edu

Humberto La Roche
Cisco Systems
USA
hlaroche@cisco.com

ABSTRACT

The demand for content such as multimedia services with high performance (e.g., ultra-low latency) requirements has increased significantly, posing heavy backhaul congestion in mobile networks. The integration of Multi-access Edge Computing (MEC) and 5G network is an emerging solution that alleviates the backhaul congestion to meet the required network performance for user equipment (UE). However, uncertainties due to user mobility cause the most challenging barrier in deciding optimal content routes from Edge Application Servers (EASs) to UEs, defined as the 5G component selection problem. To this aim, we propose a novel learning-based component selection solution for high-performance content delivery in 5G-enabled MEC that leads to minimum latency by reducing frequent handovers.

CCS CONCEPTS

• Computer systems organization → Cloud computing; **• Networks → Mobile networks**.

KEYWORDS

Multi-access Edge Computing, 5G, Content Delivery Network (CDN), Mobility, Multi-Armed Bandit.

ACM Reference Format:

Erfan Farhangi Maleki, Weibin Ma, Lena Mashayekhy, and Humberto La Roche. 2021. Learning-based Content Delivery in 5G-enabled Multi-access Edge Computing. In *SC '21: International Conference for High Performance Computing, Networking, Storage, and Analysis, November 14-19, 2021, St. Louis, MO, USA*. ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

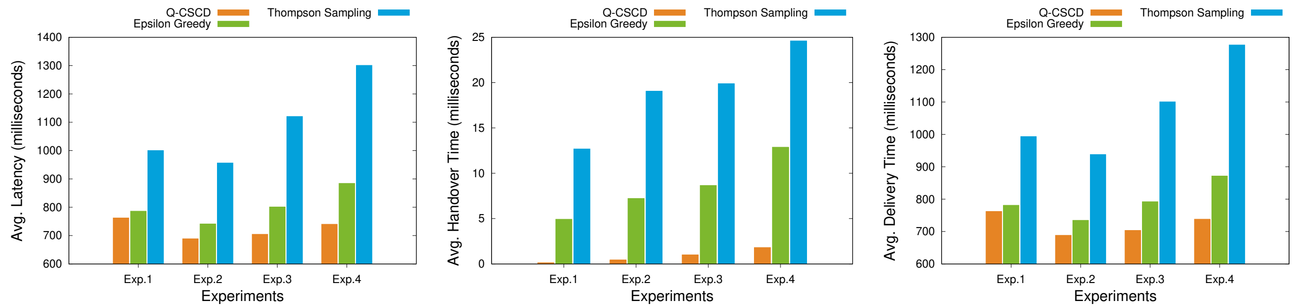
Videos will make up to 79% of the global mobile data traffic by 2022 [5], which mostly require high performance networks for content delivery to enhance the quality of experience (e.g., strict latency, service continuity) of users. Many efforts have been conducted to meet the high performance requirements of the emerging unprecedented traffic demands. In this regard, the fifth-generation (5G) network is a promising technology suitable to meet the diverse quality of service (QoS) requirements, such as ultra-low latency and ultra-high reliability [6]. However, with the rapid growth of mobile data traffic, 5G networks encounter significant challenges in capacity-limited backhaul links [10]. In consideration of this challenge, Multi-access Edge Computing (MEC) offers cloud advantages at the edge of the network close to mobile devices and delivers extremely low latency and high throughput [4, 7–9], which complements 5G. Content delivery utilizing content caching at MEC is an effective approach to better deal with the exceedingly growing traffic over the mobile networks. Moreover, it can significantly degrade duplicated transmissions of the content from backhaul links and thus improve experienced QoS for UEs. However, UEs' high mobility causes severe uncertainty in deciding the places to locate the UEs' most popular content and delivery routes.

The integration of 5G and MEC, called 5G-enabled MEC, includes the 5G components and MEC servers (EAS). The 3GPP 5G RAN architecture, which is recognized as NG-RAN, consists of radio base stations called gNBs connected to the 5G core network (5GC) and each other. The gNB component incorporates three main functional modules: the Radio Unit (RU), the Distributed Unit (DU), and the Centralized Unit (CU), which can be located separately or together in various combinations. Furthermore, each gNB must be connected to a PSA UPF to enable a UE to access the targeted content cached at an EAS. In other words, the PSA UPF plays a key role in building the content routing path between each UE and the content at an EAS. The component selection problem in 5G-enabled MEC is composed of selecting the best gNB, PSA UPF, and EAS to determine each UE's optimal routing path with minimum latency from a gNB to an EAS with the content. However, the high mobility of UEs results in multiple gNB allocations as their connected gNBs may change over time based on the UEs' new locations. This paper proposes an online learning-based approach to solve the component selection problem to minimize the content delivery latency for UEs in 5G-enabled MEC in a reasonable time.

2 SYSTEM MODEL

The system model consists of a set of UEs $\mathcal{N} = \{1, \ldots, N\}$, a set of gNBs $\mathcal{G} = \{1, \ldots, G\}$, and a set of EASs $\mathcal{M} = \{1, \ldots, M\}$. Since the

SC '21, November 14–19, 2021, St. Louis, MO, USA



(a) Total Latency **(b) Handover Time** **(c) Content Delivery Time**

Figure 1: Performance Analysis - Latency Components

PSA UPF is co-located with the EAS, PSA UPF i represents the PSA UPF co-located with EAS i. Time horizon is defined as multiple time slots shown by $\mathcal{T} = \{1, \ldots, T\}$. Allocating a new PSA UPF leads to relocating (rerouting) the previous PSA UPF to a new PSA UPF. Therefore, we define the relocation latency of a PSA UPF from EAS i to EAS j by r_{ij}. Migration latency of a content from EAS i to EAS j is denoted by s_{ij}. Each UE n requests (requires) a content of size c_n^t at time t. The transmission speeds between different components of a 5G network are denoted by: d_{ng}^t, d_{gd}, d_{gc}, d_{gi}, and d_{ij}. The objective of the 5G component selection problem is formulated as follows:

$$\text{Minimize } \mathbf{D} = \sum_{t\in\mathcal{T}} \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{M}} \sum_{j\in\mathcal{M}} (p_{nij}^t r_{ij} + q_{nij}^t s_{ij}) + \tag{1}$$

$$\sum_{t\in\mathcal{T}} \sum_{n\in\mathcal{N}} \sum_{g\in\mathcal{G}} \sum_{i\in\mathcal{M}} \sum_{j\in\mathcal{M}} k_{ngij}^t c_n^t \left(\frac{1}{d_{ng}^t} + \frac{1}{d_{gd}} + \frac{1}{d_{gc}} + \frac{1}{d_{gi}} + \frac{1}{d_{ij}}\right)$$

where p_{nij}^t, q_{nij}^t, k_{ngij}^t are the decision variables representing the rerouting of PSA UPF from EAS i to EAS j for UE n at time slot t, migration of content from EAS i to EAS j for UE n at time slot t, and the content routing path consists of gNB g, PSA UPF i, EAS j for UE n at time slot t, respectively.

3 QOS-AWARE 5G COMPONENT SELECTION FOR CONTENT DELIVERY

We design a QoS-Aware 5G Component Selection for Content Delivery (Q-CSCD) to solve the component selection problem efficiently, achieving a bounded performance. Q-CSCD learns the optimal component selection online for UEs, including gNB, PSA UPF, and EAS, to minimize latency for content delivery over time. Q-CSCD is formulated as the Multi-Armed Bandit problem. In this regard, Q-CSCD chooses an unselected routing path for every UE in its proximity at each time slot and observes its average latency per unit of the received content. During time slots with no new unselected routing path for the UE, Q-CSCD selects a routing path that has resulted in the minimum latency thus far for that UE. Q-CSCD also lowers the complexity of learning by considering just routes that could deliver the UE's requested content and are in its closer proximity. Q-CSCD applies a tradeoff between exploration and exploitation, where for each UE, it explores different routing paths to learn a range of possible new latencies. At the same time, it gives a chance to exploit a priori known optimal routing paths. Q-CSCD is adopted from the UCB1 strategy, which sets a balance between exploration and exploitation [2, 3]. However, due to the online learning characteristics of Q-CSCD, it is inevitable to select suboptimal paths and to have higher radio handover and content migrations between EASs, especially at the beginning of the learning process.

4 EXPERIMENTAL RESULTS

We compare the performance of Q-CSCD with the following approaches: 1) Epsilon Greedy: this approach balances exploration and exploitation by randomly choosing between exploration and exploitation, and 2) Thompson Sampling: this approach uses Beta distribution to prioritize choosing between paths. Our optimization criterion is to minimize the experienced latency time of users, composed of content delivery time and handover time (relocation time and migration time). The experimental setup consists of a a set of UEs scattered uniformly, while all components of a 5G network are randomly scattered over the area according to a uniform distribution. Furthermore, we utilize ThousandEyes [1] to model transmission speeds between different links in the 5G network. We consider four different-size scenarios (Exp. 1 to Exp. 4) for the experiments with the following configurations: (50 UEs, 7 gNBs, 2 PSA UPFs, and 2 EASs), (100 UEs, 8 gNBs, 3 PSA UPFs, and 3 EASs), (150 UEs, 8 gNBs, 4 PSA UPFs, and 4 EASs), and (200 UEs, 10 gNBs, 5 PSA UPFs, and 5 EASs), respectively. The average latency, handover time, and content delivery time for users per time slot are presented in Fig. 1a, Fig. 1b, and Fig. 1c, respectively. The results show Q-CSCD obtains significantly lower latency than other learning-based approaches in all experiments.

5 CONCLUSION

We tackled the component selection problem for high performance content delivery, which is a critical problem in 5G-enabled Multi-access Edge Computing. We proposed a multi-armed bandit approach, called Q-CSCD, to learn the optimal 5G components that leads to minimum experienced latency for UEs. As future work, we plan to analyze the regret of our approach and investigate the impact of different parameters on the results. We also plan to conduct more intensive experiments to compare Q-CSCD with other recent learning-based and prediction-based approaches.

ACKNOWLEDGMENT

This research was supported by Cisco grant CG#1935382.

REFERENCES

[1] 2021. Thousandeyes. Available: https://www.thousandeyes.com/.

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2 (2002), 235–256.

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* 32, 1 (2002), 48–77.

[4] Dixit Bhatta and Lena Mashayekhy. 2021. A Bifactor Approximation Algorithm for Cloudlet Placement in Edge Computing. *IEEE Transactions on Parallel and Distributed Systems (TPDS) (in press)* (2021), 1–12.

[5] GMDT Forecast. 2019. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. *Update* 2017 (2019), 2022.

[6] Najmul Hassan, Kok-Lim Alvin Yau, and Celimuge Wu. 2019. Edge computing in 5G: A review. *IEEE Access* 7 (2019), 127276–127289.

[7] Yun Chao Hu, Milan Patel, Dario Sabella, Nurit Sprecher, and Valerie Young. 2015. Mobile edge computing—A key technology towards 5G. *ETSI white paper* 11, 11 (2015), 1–16.

[8] Weibin Ma and Lena Mashayekhy. 2021. Quality-Aware Video Offloading in Mobile Edge Computing: A Data-driven Two-stage Stochastic Optimization. In *Proc. of the IEEE International Conference on Cloud Computing (IEEE CLOUD)*. IEEE, 1–6.

[9] Erfan Farhangi Maleki, Lena Mashayekhy, and Seyed Morteza Nabavinejad. 2021. Mobility-Aware Computation Offloading in Edge Computing using Machine Learning. *IEEE Transactions on Mobile Computing (TMC) (in press)* (2021), 1–12.

[10] Bhaskar Prasad Rimal, Dung Pham Van, and Martin Maier. 2017. Mobile edge computing empowered fiber-wireless access networks in the 5G era. *IEEE Communications Magazine* 55, 2 (2017), 192–200.