Efficient Offloading Task Minimization on Edge Servers Using the kNN Algorithm in Cellular Networks

Na Yeon Kim and Seong Ho Jeong
Hankuk University of Foreign Studies
nayeon718@naver.com, shjeong@hufs.ac.kr

셀룰러 네트워크에서 kNN 알고리즘을 사용한 효율적인 엣지에 작업 오프로딩 최소화

김나연, 정성호
한국외국어대학교 정보통신공학과

Abstract

This paper addresses the critical need for fast and accurate task offloading on edge servers within cellular networks, particularly in scenarios involving user mobility. Our proposed approach focuses on making task offloading decisions with minimal information. Leveraging the k-Nearest Neighbors (kNN) algorithm, we aim to expedite computation time by identifying nearby edge servers for mobile devices. Through an analysis of latency and resource utilization metrics across selected edge servers, we determine the most suitable server with the lowest latency and resource usage. By executing tasks on these optimized edge servers, it is possible to reduce task execution time, ensuring that overall tasks are completed within user expected timeframes.

Ⅰ. Introduction

Most existing studies on mobility issues assume the availability of full system information such as Mobile Edge Computing (MEC) or Base Station (BS) side information [1]. However, our paper shifts the focus on determining the optimal edge for task offloading. In this paper, we consider two crucial factors, delay and capacity, which represent both the primary advantage and challenge within edge computing frameworks.

Ⅱ. Method

The main objective of the proposed approach is to find the most suitable server with the lowest latency and resource usage. In this paper, we utilize predefined parameters in EdgeCloudSim that is an open-source simulation tool and supports cloud and edge (i.e., MEC servers deployed at BS) scenarios with mobility considerations. Specifically, we utilize the types of hardware resources and locations within this simulation framework. There are four distinct latency types: queuing, processing, transmission, and propagation delay. Table 1 illustrates how each delay is influenced by specific parameters.

Table 1. Parameters related to latency

Latency or Delay	Parameters
Propagation delay	Distance between edge and mobile device and BW
Transmission delay	Data length and transmission rate
Queuing delay	Queue depth
Processing delay	Data processing rate of edge

The user offloads the task to the nearest edge server. Therefore, the distance between them is initially analyzed. Figure 1 illustrates the five nearest edges for each user using the kNN algorithm. If there are *k* edges represented by blue dots, we concentrate solely on half of *k* or (*k*-1)/2 nodes to streamline the computation involved in identifying the suitable edges. The subsequent step involves examining the parameters of those edges depicted in Figure 1. Greater bandwidth, shorter data length, faster transmission rate, reduced queue backlog, higher processing rate contribute to lower total delay. Additionally, resource utilization, encompassing CPU, RAM, and storage, determines the extent to which an edge's capacity is employed. The total cost is derived from the summation of delay and resource utilization. The main objective is to identify the edge with the lowest associated cost, thereby optimizing task offloading decisions in cellular networks.

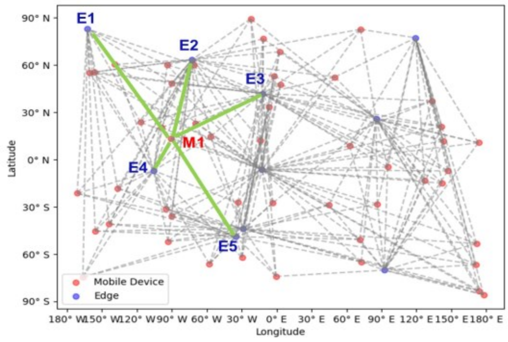

Figure 1. 3-Nearest Edges of Mobile Device

Ⅲ. Conclusion

In conclusion, incorporating network and resource utilization parameters into task offloading strategies ensures practicality and alignment with real-world challenges, ultimately enhancing system efficiency and user experience in cellular networks.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00484).

REFERENCES

[1] Ramesh S., Radhika S., Suchetana C., "A survey of mobility-aware Multi-access Edge Computing: Challenges, use cases and future directions", Ad Hoc Networks Volume 140, 2023.