Digital Twin-Assisted Efficient Offloading Model for NFV-enabled Mobile Edge Computing

Hoa Tran-Dang, Dong-Seong Kim

***Abstract*—This short paper presents a novel architecture for energy-efficient offloading in the network function virtualization (NFV)-enabled mobile edge computing (MEC) systems using digital twin (DT). The NFV technology is integrated in the server site, allowing each server to process various types of tasks. The DT is used to estimate the status information of physical objects including user equipment (UE) and edge servers, and supports the global optimization decision-making. Based on these, we finally formulate the optimization problem that minimize the total energy consumption of system while guaranteeing the satisfaction of QoS (i.e., deadlines). This framework potentially opens many research directions regarding techniques to solve the optimization problem and evaluation analysis of DT impact on the system performance.**

***Index Terms*—Mobile Edge Computing, Digital Twin, Network Function Virtualization, Radio Access Network.**

I. Introduction

Recently, mobile edge computing (MEC) has emerged as a promising ubiquitous and pervasive computing solution for IoT-connected user equipment (UE) to fulfill the requirements of 5G mobile networks. Fundamentally, MEC supports to provide the low-latency computing services by offloading computation tasks to the nearby edge servers [1]. In addition, when network function virtualization (NFV) is deployed in the server site, the executing edge servers are able to process various types of tasks. In other words, there is no restriction on offloading a task to a predetermined server [2].

However, designing an optimal task offloading is still challenging in the large-scale MEC system with heterogeneity of UEs and edge servers mainly due to the network size and dynamics. In recent years, both the academia and industry have shown great interest in developing and applying digital twin (DT) technology for intelligent resource allocation and network management in the MEC systems [3]. By integrating DT in the MEC systems, the global network status information can be monitored and estimated, thus allowing the task offloading decision to be made in the centralized manner.

There are existing related works proposing the optimal offloading in the MEC systems with the assistance of DT. For example, the DT is used to provide the estimated states of edge servers and training data to the centralized base station to derive the optimal offloading solution that minimize the offloading latency while maintaining the acceptable long-term migration cost. In [4], the DT supports UEs to select high-quality MEC servers, thus the offloading is efficient in term of energy consumption and latency. Similarly, the optimal server selection is derived by the assistance of DT in [5] through a set of iterative optimization processes.

Different from the aforementioned works, we examine a novel architecture and scenario in this short paper. First, by integrating the NFV-enabled servers, the selection of optimal server for each task is closely related to the selection of optimal routing path from UE generating the task to server. Second, we consider partial offloading mode in which a portion of task is processed locally and the rest is offloaded by multiple edge servers; thus imposing optimal offloading factor problem. Third, we evaluate the impact of DT on the efficiency of offloading performance in terms of energy consumption and QoS satisfaction.

Hoa Tran-Dang and Dong-Seong Kim are with department of IT Convergence Engineering, Kumoh National Institute of Technology, Korea, e-mail: {hoa.tran-dang, dskim}@kumoh.ac.kr.

II. System Model

A. NFV-based MEC

We consider a MEC system in which a set $\mathbb{U}$ of X UEs ($\mathbb{U} = \{UE_1, ..., UE_X\}$) request the offloading services to a set $\mathbb{E}$ of Z MEC servers ($\mathbb{E} = \{N_1, ..., N_Z\}$) through a radio access network (RAN) as shown in Fig. 1.

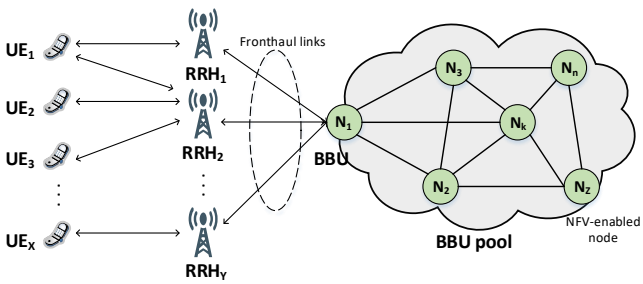

Fig. 1: NFV-enabled MEC system

In addition, the centralized RAN includes a baseband unit (BBU), which connects to a set $\mathbb{R}$ of Y RRHs (Remote Radio Heads) ($\mathbb{R} = RRH_1, ..., RRH_Y$) through a fronthaul link. The BBU is deployed in a BBU pool that is a inter-connected network of Z NFV-enabled nodes. Each node is comprised of an execution server and a routing device. Each RRH equipped with M antenna can serve a set $\mathbb{K}_u$ of K_u multiple EUs simultaneously.

B. Offloading Model in NFV-based MEC

Assuming that at the beginning of each time slot, each UE_i generates a single task T_i represented by a tuple $T_i = \langle A_i, B_i, D_i \rangle$, where A_i is the data size of task, B_i is required

CPU cycles to process the tasks, and D_i is the deadline of task execution. We consider partial offloading in which a part of task is processed locally and the rest is offloaded by multiple MEC servers. Denote α_i and β_{ij} as the portions of task T_i which are processed locally by UE i and offloaded by an edge server N_j. We have $A_i = \alpha_i A_i + \sum_{j\in\mathbb{N}} \beta_{ij} A_i$ and $B_i = \alpha_i B_i + \sum_{j\in\mathbb{N}} \beta_{ij} B_i$, where $\alpha_i + \sum_{j\in\mathbb{N}} \beta_{ij} = 1$ and $0 \le \alpha_i \le 1$ and $0 \le \beta_{ij} \le 1$.

C. DT Model

We consider the DT models for the UEs and MEC servers. Define $\mathbb{D}_i^u = \{f_i^u, \overline{f}_i^u\}$ and $\mathbb{D}_j^e = \{f_j^e, \overline{f}_j^e\}$ as the DT models for an UE *i* and an edge server N_j, where f_i^u and f_j^e are CPU frequency estimated by the DT models and $\overline{f}_i^u$, $\overline{f}_j^e$ are the deviation between the real value and the estimation. The DT models can be deployed in the power-rich resource nodes such as an edger server or the BBU as the centralized management, thus it can derive the global decision making.

D. Offloading Delay Model

The radio transmission delay of task T_i as transmitting the data from UE_i to RRH_k is $T_{ik}^{tx} = (1-\alpha_i)A_i/r_{ik}$, where r_{ik} is the achievable data rate and $r_{ik} = BW \log_2(1 + SINR_i)$. BW is the bandwidth of RAN and $SINR_i$, the signal to interference plus noise ratio of UE_i, can be achieved in advanced [6].

$$SINR_i = \frac{\|h_{k,i}\|^2 p_i}{\sum_{j\in\mathbb{U}\setminus\{i\}} \frac{|h_{k,i}^H h_{k,j}|^2}{\|h_{k,i}\|^2} p_j + \sigma_n^2}, \forall i \in \mathbb{K}_k \quad (1)$$

In this formulation, $h_{k,i}$ is the channel vector between UE_i and RRH_k. The sum of data rates of UEs served by RRH_k must be less than the capacity of its fronthaul link, i.e., $\sum_{i\in\mathbb{K}_k} r_{ik} \le B_{f;k}, \forall k \in \mathbb{R}$.

Let $C_i = \gamma B_i$ (in cycles) denote the required computation resource, where γ is the complexity of the tasks in cycles/bit. The estimated time required to execute the portion α_i of task T_i locally is computed as $\overline{T}_i^{lc} = \frac{\alpha_i C_i}{f_i}$. The computing delay gap between the real value and DT estimation is given by $\delta T_i^{lc} = \frac{\alpha_i C_i \overline{f}_i}{f_i(f_i - \overline{f}_i)}$. Consequently, the actual time for local computing at UE_i can be achieved by $T_i^{lc} = \delta T_i^{lc} + \overline{T}_i^{lc}$.

The estimated latency of edge server N_k to execute the portion β_{ik} of task T_i is given by $\overline{T}_{ik}^{e} = \frac{\beta_{ik} C_i}{f_k}$. The latency gap between real value and DT estimation is calculated as $\delta T_{ik}^{e} = \frac{\beta_{ik} C_i \overline{f}_k}{f_k(f_k - \overline{f}_k)}$. The actual latency for executing the portion β_{ik} of T_i at the edge DT can be given by $T_{ik}^{e} = \delta T_i^{e} + \overline{T}_i^{e}$.

Finally, for the task T_i, the total DT latency in the system can be expressed by: $T_i^{tot} = T_i^{lc} + \max_{k\in\mathbb{R}} T_{ik}^{tx} + \max_{k\in\mathbb{N}} T_{ik}^{e}$.

III. Problem Formulation

The total energy consumption of UE_i including the local computation energy (E_i^{lc}) and transmission energy (E_i^{tx}) is given by $E_i = E_i^{lc} + E_i^{tx} = \alpha_i \frac{\theta}{2} B_i (f_i - \overline{f}_i)^2 + \sum_{k\in\mathbb{K}} p_i \frac{\beta_{ik} B_i}{r_{ik}}$, where $\theta/2$ is a constant expressing the average switched capacitance and the average activity factor of UE_i [7]. The QoS constraint is to ensure that the total latency of task is must be less than the deadline, i.e., $T_i^{tot} \le D_i, \forall i \in \mathbb{U}$. The objective of system is to minimize the total energy consumption of UEs based on optimizing offloading policies (α_i, β_{ij}), edge server selection (β_{ij}), transmit power (p_i), and estimated CPU frequency of the UEs and MEC servers (f_i^u, $\forall i \in \mathbb{U}$, f_k^e, $\forall k \in \mathbb{N}$).

IV. Conclusions and Future Works

This paper introduced the computation offloading model assisted by the digital twin in the NFV-enabled MEC system. DT allows the central base station estimating the status of nodes in the network, just facilitating the modeling and optimizing the resource allocation and management. Based on these, we proposed the optimization framework to jointly optimizing the transmit power of UEs, offloading policies, and edge server selection to offload the computation tasks.

The proposed offloading model potentially exposes many directions for future research works. First of all, the optimization problem is non-convexity. Therefore, to solve it requires developing techniques to transform the problem into convex-like form such as decomposing the problem into subproblems [6] or iterative methods [5]. Secondly, the simulation environment should be modeled to evaluate the performance of systems under the impact of DT technology.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2020-2020-0-01612) and supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2018R1A6A1A03024003), and Korea Research Fellowship Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2020R1I1A1A01073019).

References

[1] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Communications Surveys and Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[2] B. Yi, X. Wang, K. Li, S. k. Das, and M. Huang, "A comprehensive survey of network function virtualization," *Computer Networks*, vol. 133, pp. 212–262, 2018.

[3] W. Sun, H. Zhang, R. Wang, and Y. Zhang, "Reducing offloading latency for digital twin edge networks in 6g," vol. 69, no. 10, pp. 12 240–12 251.

[4] T. Liu, L. Tang, W. Wang, Q. Chen, and X. Zeng, "Digital-twin-assisted task offloading based on edge collaboration in the digital twin edge network," *IEEE Internet of Things Journal*, vol. 9, no. 2, pp. 1427–1444.

[5] T. Do-Duy, D. Van Huynh, O. A. Dobre, B. Canberk, and T. Q. Duong, "Digital twin-aided intelligent offloading with edge selection in mobile edge computing," *IEEE Wireless Communications Letters*, vol. 11, no. 4, pp. 806–810.

[6] M. Tajallifar, S. Ebrahimi, M. R. Javan, N. Mokari, and L. Chiaraviglio, "Energy-efficient task offloading under e2e latency constraints," *IEEE Transactions on Communications*, vol. 70, no. 3, pp. 1711–1725.

[7] W. Sun, P. Wang, N. Xu, G. Wang, and Y. Zhang, "Dynamic digital twin and distributed incentives for resource allocation in aerial-assisted internet of vehicles," *IEEE Internet of Things Journal*, vol. 9, no. 8, pp. 5839–5852.